RESEARCH ARTICLE

REVISED Polysomal mRNA Association and Gene Expression in *Trypanosoma brucei* [version 3; peer review: 3 approved, 1 approved with reservations]

Michele Tinti*, Anna Kelner-Mirôn*, Lizzie J. Marriott, Michael A.J. Ferguson

Wellcome Centre for Anti-Infectives Research (WCAIR), School of Life Sciences, University of Dundee, Dundee, Dundee, UK

* Equal contributors

v3 **First published:** 22 Feb 2021, **6**:36
https://doi.org/10.12688/wellcomeopenres.16430.1

Second version: 26 Aug 2021, **6**:36
https://doi.org/10.12688/wellcomeopenres.16430.2

Latest published: 01 Feb 2022, **6**:36
https://doi.org/10.12688/wellcomeopenres.16430.3

Abstract

Background: The contrasting physiological environments of *Trypanosoma brucei* procyclic (insect vector) and bloodstream (mammalian host) forms necessitates deployment of different molecular processes and, therefore, changes in protein expression. Transcriptional regulation is unusual in *T. brucei* because the arrangement of genes is polycistronic; however, genes which are transcribed together are subsequently cleaved into separate mRNAs by *trans*-splicing. Following pre-mRNA processing, the regulation of mature mRNA stability is a tightly controlled cellular process. While many stage-specific transcripts have been identified, previous studies using RNA-seq suggest that changes in overall transcript level do not necessarily reflect the abundance of the corresponding protein.
Methods: To better understand the regulation of gene expression in *T. brucei*, we performed a bioinformatic analysis of RNA-seq on total, sub-polysomal, and polysomal mRNA samples. We further cross-referenced our dataset with a previously published proteomics dataset to identify new protein coding sequences.
Results: Our analyses showed that several long non-coding RNAs are more abundant in the sub-polysome samples, which possibly implicates them in regulating cellular differentiation in *T. brucei*. We also improved the annotation of the *T.brucei* genome by identifying new putative protein coding transcripts that were confirmed by mass spectrometry data.
Conclusions: Several long non-coding RNAs are more abundant in the sub-polysome cellular fractions and might pay a role in the regulation of gene expression. We hope that these data will be of wide general interest, as well as being of specific value to researchers studying gene regulation expression and life stage transitions in *T. brucei*.

Open Peer Review

Approval Status ? ✓ ✓ ✓

	1	2	3	4
version 3 (revision) 01 Feb 2022		✓ view		✓ view
version 2 (revision) 26 Aug 2021				
version 1 22 Feb 2021	? view	? view	✓ view	? view

1. **Shulamit Michaeli**, Bar-Ilan University, Ramat-Gan, Israel
2. **Esteban Erben**, IIBIO-UNSAM, Buenos Aires, Argentina
3. **Magdalena Radwanska**, Ghent University Global Campus, Incheon, South Korea
   Ghent University, Ghent, Belgium
4. **Osvaldo P de Melo Neto**, Instituto Aggeu Magalhães, Fiocruz, Recife, Brazil
   Antonio Rezende, Instituto Aggeu Magalhães, Fiocruz, Recife, Brazil

Any reports and responses or comments on the

Keywords
RNA-seq, mRNA, Polysome, Trypanosoma brucei, Bloodstream form, Procyclic form, machine learning

article can be found at the end of the article.

Corresponding author: Michael A.J. Ferguson (m.a.j.ferguson@dundee.ac.uk)

Author roles: Tinti M: Data Curation, Formal Analysis, Software, Visualization, Writing – Original Draft Preparation; **Kelner-Mirôn A**: Investigation, Methodology, Writing – Original Draft Preparation; **Marriott LJ**: Data Curation, Formal Analysis, Writing – Original Draft Preparation; **Ferguson MAJ**: Conceptualization, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This study was supported by the Wellcome Trust through an Investigator Award to MASF [101842, https://doi.org/10.35802/101842], which also supported MT; and a PhD studentship to AKM [093712, https://doi.org/10.35802/093712].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Tinti M, Kelner-Mirôn A, Marriott LJ and Ferguson MAJ. **Polysomal mRNA Association and Gene Expression in *Trypanosoma brucei* [version 3; peer review: 3 approved, 1 approved with reservations]** Wellcome Open Research 2022, **6**:36 https://doi.org/10.12688/wellcomeopenres.16430.3

First published: 22 Feb 2021, **6**:36 https://doi.org/10.12688/wellcomeopenres.16430.1

REVISED **Amendments from Version 2**

The new version of the manuscript provides a new chapter (Expression Analysis of lncRNAs and surrounding genes) containing a more robust statistical analysis to correlate the expression of the lncRNAs and the genes at their 5′ or 3′. We used only reads mapped to the transcript coding sequences and to the long non-coding RNAs to provide this analysis, with the underlining data deposited in a new repository. Also, the visualization of our analysis is provided in the new Figure 17. We updated Figure 5 to correct a normalisation error and we updated the discussion to reflect comments made by the referees.

Any further responses from the reviewers can be found at the end of the article

Introduction

Trypanosoma brucei, a protozoan parasite transmitted by the tsetse fly, causes human African trypanosomiasis (HAT) and nagana in cattle[1]. The parasite undergoes a complex lifecycle between its insect vectors and mammalian hosts[2]: Slender bloodstream form (BSF) parasite proliferate predominantly in the blood and lymph of the infected mammalian host in the first stage of the disease and the second neurological stage of the disease occurs when these parasites cross the blood-brain barrier. Some of the slender BSF parasites differentiate into non-replicative stumpy forms in the bloodstream and these are pre-adapted for transformation into replicating procyclic form in the testse vector midgut. Procyclic forms further differentiate into replicating epimastigote and then non-dividing metacyclic trypomastigote forms during parasite migration to the tsetse salivary glands. The metacyclic parasites are transferred to a new host during a bloodmeal and after differentiation into slender BSF parasites, the lifecycle in complete. The BSF and PCF parasites are the easiest to propagate in the laboratory and are the most studied.

Transcription is particularly interesting in *T. brucei* because the arrangement of its genes is polycistronic. Thus, RNA Polymerase II (RNA Pol II) transcribes protein-coding genes into large polycistrons containing several transcripts. However, the polycistron does not linger as it is co-transcriptionally processed into individual mRNAs[3]. The processing of the transcription unit occurs by trans-splicing coupled to cleavage of the 3´ end by the polyadenylation machinery for poly(A) addition[4,5]. During trans-splicing, a capped 39-nucleotide (nt) spliced leader (SL) mini-exon is added to the 5′ termini of mRNAs. The SL sequence was first discovered when two different VSG transcripts were found with an identical leader sequence at their 5′ ends, which was not evident in their genomic sequence[6–8]. This mini-exon is independently transcribed from a tandem array of 140-nt spliced leader (SL) RNA genes[9,10].

Recent studies using RNA-seq have greatly improved our understanding of the *T. brucei* transcriptional landscape across the BSF and PCF life stages[2,11–15]. These studies have found new transcripts, many non-coding RNAs, and facilitated the correction of numerous annotations across the *T. brucei* genome. While several aspects of translational control have been investigated in *T. brucei*, there are only a few examples of polysome profile analysis that have explored the efficiency of translation between BSF and PCF parasites[12,16]. Numerous 80S ribosomes can be translating an mRNA transcript at the same time, producing so-called 'polysomes'[17]. The number of ribosomes on an mRNA generally reflects that transcript's rate of translation under given conditions[18]. Further, a particular mRNA's higher or lower than average association with ribosomes indicates the potential involvement of gene-specific regulatory mechanisms[19].

To make a contribution to our understanding of the regulation of gene expression in trypanosomes, we investigated mRNA recruitment to ribosomes with RNA-seq of total polyA+, sub-polysomal, and poly-ribosomal mRNA purified from BSF and PCF life stages of *T. brucei*.

Methods

Cell culture

T. brucei bloodstream form cells, Lister strain 427, VSG variant MITat1.2[20] (kindly provided by Prof. George Cross) were cultured at 37°C with 5% CO_2 in cell culture flasks with filter lids (Greiner). Cells were grown to a maximum density of $3x10^6$ cells/ml in HMI-9T medium (HMI-9 powder, Gibco Catalog Number: 07490915N). HMI-9T contains variations on the HMI-9 medium described in 21: thioglycerol (Sigma, Catalog Number: m6145) was used instead of β-mercaptoethanol, and GlutaMAX (Gibco, Catalog Number: 35050-38) was used instead of L-glutamine for their increased stability. *T. brucei* procyclic form transgenic cell line 29.13.6, Lister strain 427 (kindly provided by Prof. George Cross) was cultured at 28°C in Becton Dickinson culture flasks. Cells were grown to a maximum density of $4x10^7$ cells/ml in SDM-79 medium (Invitrogen, custom made on request, Catalog Number: N/A)[22] supplemented with 15% fetal bovine serum (FBS) (PAA, Catalog Number: A11-101), GlutaMAX (Gibco, Catalog Number: 35050-38), and 15 µg/ml hemin (Sigma, Catalog Number: H9039).

Polysome fractionation and RNA extraction

Log-phase cultures of *T. brucei* BSF and PCF cells were incubated with 50 µg/ml cycloheximide (Sigma, Catalog Number: C4859) for 10 min prior to the start of polysome purification procedures. Cells were pelleted by centrifugation at 800 g for 10 min at 4°C. PCF cells were washed with PBS (137 mM NaCl, VWR Catalog Number: X190; 2.7 mM KCl, VWR Catalog Number: ICNA0215194401; 10 mM Na_2HPO_4, VWR Catalog Number: 4062-01; 2 mM KH_2PO_4 pH 7.4, VWR Catalog Number: 26925.295) containing 1 mg/ml cycloheximide (Sigma Catalog Number: C4859), while BSF cells were washed with trypanosome dilution buffer (5 mM KCl, VWR Catalog Number: ICNA0215194401; 80 mM NaCl, VWR Catalog Number: X190; 1 mM $MgSO_4$ VWR Catalog Number:

2506-01; 20 mM Na_2HPO_4, VWR Catalog Number: 4062-01; 2 mM NaH_2PO_4, VWR Catalog Number: ICNA0219550091; 20 mM glucose pH 7.4, VWR Catalog Number: 1910-05) containing 1 mg/ml cycloheximide (Sigma Catalog Number: C4859). Cells were resuspended in polysome lysis buffer (120 mM KCl, VWR Catalog Number: ICNA0215194401; 2 mM $MgCl_2$ VWR Catalog Number: ICNA0520984480; 20 mM Tris-HCl pH 7.5 VWR Catalog Number: ICNA04816100; 1 mM DTT Sigma Catalog Number: 10708984001; 1% n-octylglycoside Sigma Catalog Number: 10634425001; 50 µl RNAsin Promega Catalog Number: N2111; 2 µg/ml leupeptin Sigma Catalog Number: L2884; 1 µg/ml aprotinin Sigma Catalog Number: A6279; 1 µM TLCK Sigma Catalog Number: 90182; 1 mM PMSF Sigma Catalog Number: 10837091001; 1mg/ml cycloheximide Sigma Catalog Number: C4859). The detergent n-octylglycoside (NOG) was chosen because it does not absorb at 254 nm. The lysates were loaded on top of 10 ml sucrose (Sigma Catalog Number: S0389) gradients (5 increments, 2ml each: 10%–50% sucrose) and centrifuged for 2 h at 38,000 rpm at 4°C in a Beckman ultracentrifuge using a SW41Ti rotor. Gradients were fractionated (0.5 ml fractions) and analysed for nucleic acid content by a Nanodrop spectrophotometer at 254 nm. RNA was purified using RNeasy kits (Qiagen, Catalog Number: 74104) from pooled sub-polysome and poly-ribosomal fractions. Gradient analysis was also performed using a gradient collector (Teledyne) with continuous monitoring at 254 nm. Individual fractions were collected with a Foxy Jr. (Teledyne) fraction collector. Following collection, the RNA from each sample was purified as above and pooled according to the sub-polysomal and polysomal fractions identified in the absorbance trace.

Total RNA was extracted from bloodstream and procyclic form *T. brucei* using the RNeasy Mini Extraction Kit (Qiagen, Catalog Number: 74104). The protocol was carried out according to manufacturer's instructions with a few deviations for *T. brucei*. Cells were centrifuged for 10 min, 800 x *g* at room temperature, media was aspirated and the cell pellets were resuspended in buffer RLT (Qiagen, Catalog Number: 79216) and β-mercaptoethanol (Sigma Catalog Number: 444203) was added at a 1:100 dilution. One volume of 70% ethanol (Sigma Catalog Number: 51976) was added to the lysate and the mixture was transferred to the provided column. RNA was bound to the column by centrifugation for 15 sec, 10,000 x *g*. The column was then washed with Buffer RWI and twice with Buffer RPE (Qiagen, Catalog Number: 1018013). Following the washes, the column was transferred to a sterile (RNAse free) Eppendorf tube (Thermofisher, Catalog Number: AM12400), and the RNA was eluted in 50 µl RNase-free H_2O (Thermofisher, Catalog Number: AM9916). The RNA concentration was then estimated from the A_{260} value using a Nanodrop 2000c spectrophotometer (Thermo) with path length settings adjusted for RNA (40). Following quantitation, the purified RNA was subsequently used for RNA-seq cDNA library preparation.

Preparation of cDNA libraries for RNA-seq

Total RNA, sub-polysomal, and poly-ribosomal RNA was isolated from BSF and PCF *T. brucei* followed by poly(A) mRNA enrichment with poly-T oligomers attached magnetic beads (Illumina). The mRNA was then fragmented into 200 nt fragments using Covaris Adaptive Focused Acoustics process with the following operating conditions: Sample volume 130 µl, duty cycle 10%, intensity 5, cycles per burst 200, processing time 60 s, water bath temperature 4°C, power mode frequency sweeping, degassing mode continuous. Fragmented mRNA was concentrated by ethanol precipitation and measured on an RNA Pico chip (Agilent 2100 Bioanalyzer). The first strand of cDNA was synthesized using reverse transcriptase (Invitrogen Life Technologies, Catalog Number: 18064-022) and random primers (Invitrogen Life Technologies, Catalog Number: 1880007) using a Omnigene thermal cycler (25°C for 10 min. 42°C for 50 min, 70°C for 15 min), followed by second strand cDNA synthesis using a Omnigene thermal cycler (16°C for 60 min), producing double-stranded cDNA (NEBNext mRNA library kit for Illumina, NEB, Catalog Number: E6100. To blunt-end the DNA fragments, an end repair reaction was performed with Klenow polymerase (NEB, Catalog Number: M0210L), T4 DNA polymerase (NEB, Catalog Number: M0203L), and T4 polynucleotide kinase (NEB, Catalog Number: M0201L). A single 3´ adenosine overhang was added to the cDNA allowing the ligation of Illumina adaptors. These adaptors contain primer sites both for sequencing and complimentary annealing onto the Illumina flow cell surface (Top adapter: 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3' Bottom adapter 5′-GATCGGAAGAGCGGTTCAGCAGGAATGCCG AG-3'). Adaptor ligated cDNA fragments were measured on an Agilent DNA chip. The final cDNA libraries were sequenced on a HiSeq2000 (Illumina).

Bioinformatic analysis

The software versions of the packages used for the bioinformatic analysis are listed in the file named "package_versions.txt" and deposited in the zenodo repository mtinti/polysome. The FASTQ files of technical replicates were concatenated together. The forward and reverse paired-end reads of the biological replicates (B_tot: 1 to 3, B_pol: 1 to 3, B_sub: 1 to 3, P_tot: 1 to 3, P_pol: 1 to 3, P_sub: 1 to 3, where B=BSF, P=PCF, tot=total, pol=polysomal, sub=sub-polysomal) were aligned to the reference genome v46 of *T. brucei* clone TREU927 and 427_2018 downloaded from TriTrypDB[23] using Bowtie2[24], with the 'very-sensitive-local' pre-set alignment option. The alignments were converted to BAM format, reference sorted and indexed with SAMtools[25]. The genome coverage of the aligned reads was extracted from the BAM files using bedtools[26] with the -bg option to output bedGraph files. Fragment counts were determined from the BAM files using featureCounts[27] with parameters: -p (pair end) -B (both ends successfully aligned) -C (skip fragments that have their two ends aligned to different chromosome) -M (count multi-mapping) -O (match overlapping features) -t transcript (count level) -g gene_id (summarization level).

Assembly of Poly A and Spliced Leader Tracks

Alignments with properly paired reads were extracted with SAMtool view using the -f 2 option and parsed with a custom python script to extract the paired reads containing the last 14 bases of the spliced leader sequence (GTGAGGCCTCGCGA) in forward or reverse complement orientation. We used the last 14 bases as they are unique[28]. The same script was used to extract reads containing poly(A) tracts of at least 10 bases that are often found at the intergenic regions of *T. brucei*[29]. The aligned reads were saved in BAM format and used to create genomic track coverage in bedGraph format.

Assembling *T. brucei* transcripts

The GFF annotation file for v46 of *T. brucei* clone TREU927 was downloaded from TriTrypDB and converted to GTF with gffread[30]. The gene annotation file was supplemented with a recent prediction of long non-coding RNAs[31] (doi: https://doi.org/10.1101/2020.05.03.074625). Hypothetical new transcripts were predicted using Trinity[32] and Scallop[33]. First, we identified new predicted genes with Scallop that was run for each biological replicate. The scallop predictions in GTF format were filtered to include only genes in intergenic regions that did not have any overlap with previously annotated genes. To achieve this, the GTF prediction files and the GTF reference file were converted to bed format with gtf2bed and intersected using bedops[34]. The filtered regions were converted back to GTF format, merged in a set of unique prediction with StringTie[35] and added to the reference GTF file. In a second run, we used Trinity that was executed with the genome guided and jaccard clip parameters for each biological replicate. The predicted Trinity gene sequences were aligned to the TREU927 genome with gmap[36] and the GFF output files of gmap were converted to GTF with gffread[30]. From this point, the same filtering methods used for the Scallop predictions were applied to the Trinity predictions that were added to the reference GTF file. Both Trinity and Scallop were used as they were found to identify different sets of transcribed regions. However, both assemblers were developed for eukaryotic genes with introns, and we struggled to apply the assemblers in *T. brucei*. Particularly Trinity was prone to assemble transcripts encompassing several genes. For this reason, we run Scallop first to annotate new transcripts in regions without any previous annotation. Then, we repeat the same analysis with Trinity, again considering only regions without previous annotation. We also downloaded from GenBank[37] the genomic sequences and GFF annotation files for the entries:M94286 (maxicircle sequences), FM162566 427 VSG bloodstream form expression site 1 (BES1) locus, FM162567 427 BES2 locus and the minicircle sequences L25588, L25589, L25590, M15321. The GFF downloaded from GenBank were converted to GTF files with Biopython[38]. We also constructed a synthetic chromosome of VSG 427 gene transcripts with the sequences deposited at http://tryps.rockefeller.edu/ using the link http://129.85.245.250/Downloads/vsgs_tb427_all_atleast-150aas_cds.txt. The VSG sequences were concatenated with random DNA sequences of 50 base pairs to produce the synthetic chromosome (named fake_vsgs) and a GTF annotation file was produced. All the GTF annotation files were concatenated together as well as the gene sequences to produce a new assembly named tb927_5 (tb927_5.gtf).

Quality control

The quality of alignments were evaluated with Qualimap2[39] using the bamqc and rnaseq options. The Qualimap2 output files, and the outputs of fastp, bowtie2, Picard Mark Duplicates, SAMtools flagastat, SAMtools stats and featureCounts were aggregated with MultiQC[40], inspected and made available at https://polysome-qc.onrender.com. Dimensionality reduction was performed with the MDS algorithm implemented in SciPy[41] after log2 transform of the read counts of the top 500 expressed gene. The length and GC content of the predicted transcripts were extracted using bedtools nuc function after converting the GTF annotation file to bed format. The GC and length content biases were assessed with the cqn package for R[42] after removing genes with low counts using edgeR[43]. FPKM values for the dataset visualization were extracted using the cqn package for R.

Dataset visualization

Zero counts were replaced by the minimum value counts column-wise. The ANOVA-like test in edgeR was used to retain genes that differ in abundance in at least one of the samples with a false discovery rate <1%.

RadViz

The RadViz function implemented in the pandas python library[44] was modified and used for the visualization. For each gene the median value of the three biological replicates was computed for each experiment (B_tot: 1 to 3, B_pol: 1 to 3, B_sub: 1 to 3, P_tot: 1 to 3, P_pol: 1 to 3, P_sub: 1 to 3). For visualization, each gene was colour coded and assigned to one of the six experiments (B_tot, B_pol, B_sub, P_tot, P_pol, P_sub) where it showed the maximum abundance value.

Clustering

The dataset was normalized raw-wise with a standard scale approach, by subtracting the minimum value and dividing by the maximum value minus the minimum value, for each gene count. The optimal number of clusters was determined with the elbow approach using the KElbowVisualizer function implemented in the yellowbrick python package[45]. The dataset was divided in 4 clusters using the K-means algorithm implemented in the scikit-learn python package[46]. The columns were clustered as well using the clustermap function implemented in seaborn.

lncRNAs enrichment

The first spreadsheet "Ksplice lncRNAs" in Supplemental Table 1 of doi: https://doi.org/10.1101/2020.05.03.074625[31] was used to extract the hypothetical long non-coding mRNAs. The hypergeometric test implemented in scipy stats[41] was used to compute the enrichment p-value for long non-coding genes in each cluster.

mRNA half-life

The "BS mRNA half-life (min)" and "PC mRNA half-life (min)" columns from Table S5 of Antwi *et al.*, 2016[12] were used to extract the mRNA half-lives. The gene IDs were converted to those of version 46 of the TREU927 genome using TryTripDB.

GO term enrichment

The GO enrichment analysis was performed with the goatools python package[47]. The go-basic.obo file was downloaded with the goatools python package. The gaf association file was downloaded from TritrypDB. Enriched go term p-values were corrected with the Bonferroni option in goatools and filtered at 1% false discovery rate. For visualization, the GO terms were further filtered to include terms appearing uniquely in one of the clusters. The enriched GO terms in each cluster were sorted according to the adjusted p-value and the top-5 GO terms retained.

Identification of new protein coding genes

The Raw files described in our protein half-lives paper[48] were processed in MaxQuant with the same parameters used to compute the iBAQ values, except that the predicted amino acidic sequences from the open reading frames downloaded from TriTrypDB version 46 were used. The start and end coordinates of the identified peptides were retrieved from the peptides.txt output files and organized in bed format. The coverage values of the genomic peptide coordinates in the bed file were set to 1. The file was sorted with the sort-bed function in bedtools. We then extracted the new gene predictions from our assembled GTF file and converted them to bed format. Subsequently, we used the bedextract function in bedtools to extract the peptides mapped to new predicted transcripts. The web interface of the phobius program[49] was used to search for transmembrane domains and the web interfece ot the signalP algoritms 3.1 and 5.0[50] were used to search for signal peptides. The blast[51] searches were performed with the web interfaces implemented at the NCBI or TriTrypDB. The Clustal Omega analysis were performed with the web interfaces implemented at EMBL-EBI[52].

Coverage Visualisation. The software versions of the packages used for the visualisation of the bedGraph files are listed in the file named "package_versions.txt" and deposited in the zenodo repository mtinti/polysome_coverage. The bedGraph files were visualized with the svist4get python package[53].

Comparison with previous work

Transcription competency. Table S5 from Antwi *et al.*, 2016[12] was downloaded and the Ribosomes/kb on polysomes values were extracted from spreadsheet 1 (PCF) and spreadsheet 2 (BSF). Gene names were mapped to the version 46 of TREU927 genome using the gene search service at TriTrypDB[23]. Fragment counts for our dataset were determined from the BAM files using featureCounts[27] with parameters: -p -B -C -M -T 8 -t CDS -f to count only reads mapped to CDS regions. The read counts were filtered for low counts and normalized using edgeR[43]. Before computing the fraction of transcripts in polysomes, the polysome read counts were multiplied by 0.7 and the sub-polysome read counts were multiplied by 0.3 to correct for the total amount of mRNA found in polysome (70%) and sub-polysomal fractions (30%)[12]. The median of the fraction of transcripts in polysomal fraction was computed for the three biological replicates of BSF and PCF life stages and compared to the values reported in Antwi *et al.*, 2016[12]. The Pearson correlation coefficients between samples were computed with the python package pandas[41].

Ribosome profile

The fastq files for the ribosome profile experiment were downloaded from the ENA archive[54] with accession number PRJEB4801 and processed in a similar way as reported in Vasquez *et al.* 2014[2]. Briefly, the fastq files for the BSF and PCF biological replicates samples were concatenated together and the Illumina adaptor sequences were trimmed with the fastp package[55]. Sequences shorter than 20 bases were removed with the fastp package[55]. Reads were aligned, counted, and normalized as described above. The aligned reads in BAM format were used to create genomic track coverage in bedGraph format.

Sub-polysome / polysome differential abundance analysis

Differential abundance analyses were carried out with edgeR using generalized linear models (GLM) and the correction factors provided by the cqn package. In this study, we tested the differential abundance between the sub-polysome and polysome samples of the BSF and PCF life stages. To study the effect of the lncRNAs on the surrounding genes, we mapped again the data using the annotations of the lncRNAs and the CDS of protein coding genes. We then created a third model to identify the transcripts with differential abundance between the sub-polysomal samples (BSF and PCF) against the polysomal samples (BSF and PCF). The p-values of the test were corrected with the topTags function in R using the Benjamini–Hochberg method.

For the McNemar's test of paired samples we counted A) the lncRNAs more abundant in the polysome fraction (log fold change > 0); B) the lncRNAs more abundant in the sub-polysomal fraction (log fold change < 0); C) The 5' genes respect to the lncRNAs more abundant in the polysome fraction (log fold change > 0); D) The 5' genes respect to the lncRNAs more abundant in the sub-polysome fraction (log fold change < 0). We than used the McNemar's test implemented in the statsmodels python package as [[A+C , A+D] , [B+C , B+D]]. The same test was performed using the lncRNAs and the genes at the 3' of the lncRNAs. Only genes and lncRNAs with an FDR <0.01% were considered. The regplot function

of the seaborn python package was used for the LOWESS fitting.

The code to reproduce the analysis pipeline and the figures, the raw data and additional python scripts used for this study are available at GitHub.

Results

In our study, cells were treated with the antibiotic cycloheximide to prevent polysome run-off during sample preparation. Cycloheximide binds to the 60S ribosomal subunit and arrests translation elongation by inhibiting release of the deacylated tRNA from the ribosome E site, thereby stalling the ribosomes on mRNA in a polysomal state[56]. The high protein content of polysomes allows them to be separated throughout a sucrose gradient according to the number of ribosomes attached to the mRNA (Figure 1). To prepare samples for RNA-seq, cDNA libraries were generated from both total mRNA, polysome-associated mRNA and sub-polysomal mRNA transcripts. It is important to note that our procedure enabled the libraries to be completed without PCR amplification, therefore eliminating sample bias associated with variable amplification. In all, three (1 to 3) biological and three technical replicates of total (tot), sub-polysomal (sub), and polysomal (pol) mRNA RNA-seq experiments were performed for BSF (B) and PCF (P) life stages.

Assembling a reference transcriptome

Whole transcriptome experiments offer valuable resources to detect new genes and improve gene models. For this reason, we decided to create a complete TREU927 transcriptome assembly before assigning our reads to the reference gene set. To this

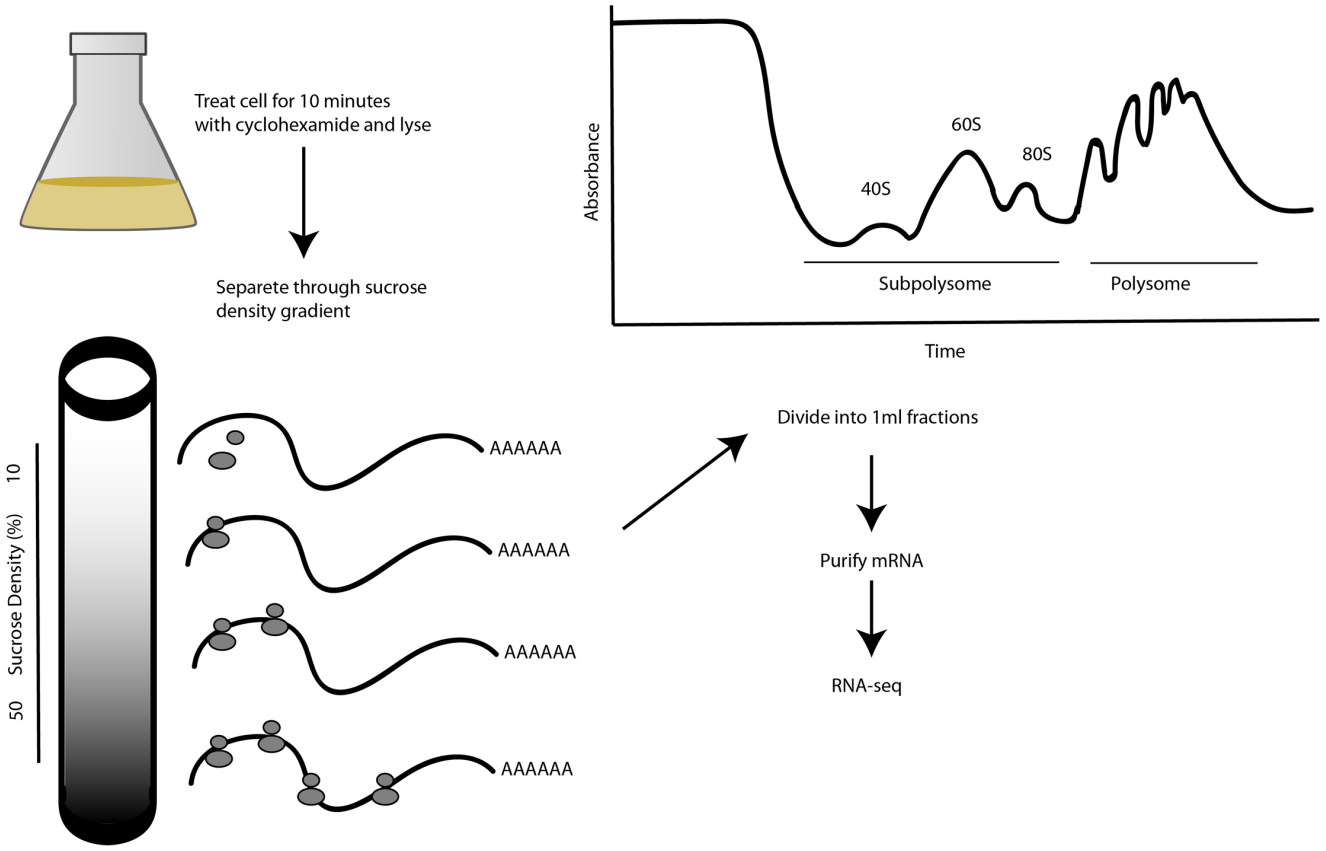

Figure 1. Experimental design. Cycloheximide-treated cells are lysed by detergent and their contents separated by centrifugation through a sucrose gradient. A representative 254 nanometer (nm) absorbance trace for nucleic acids in a Bloodstream Form (BSF) lysate density gradient is shown, normalized to the absorbance of a blank gradient. The earliest fractions contain the sub-polysomal fraction and the latest fractions contain the polysomal fraction. Free monosomes (80S) and ribosomal subunits (40S and 60S) are indicated. The messenger RNA (mRNA) transcripts from total, sub-polysomal and polysomal RNA were purified on immobilized oligo-dT for RNA sequencing (RNA-seq).

end, we first added a set of newly predicted genes described by Guegan, *et al.*[31] encoding mostly long non-coding RNA (lncRNA). Subsequently, we implemented a genome-guided approach to annotating new genes discovered from our dataset. This strategy consisted of mapping reads along the reference genome, followed by gene prediction (see Methods). This final step allowed us to extend the number of transcribed genome loci from 11725 to 15743 (an increase of 34%).

To aid the visualization of the newly predicted genes and assess the quality of the transcript boundaries, we extracted from all the samples the reads containing a spliced leader sequence and poly(A) genomic tract of >9 bases. The spliced leader sequence is present at the beginning of all mature trypanosome transcripts and can be used to determine the exact 5' boundary of the gene. The poly(A) genomic tracts are often present in intergenic regions and can help to determine the 3' gene boundaries[29]. It is useful to note that the script we used to select the poly(A) genomic tracts also selects reads with poly(A) mRNA tails. However, we did not distinguish between poly(A) mRNA tails or poly(A) genomic tracts as both are useful to define gene boundaries[57].

Quality control

The RNA-seq reads were aligned to the TREU927 reference genome, and the numbers of fragments mapping to our assembled gene list were computed. We evaluated the quality of our dataset at several levels. First, we used multidimensional scaling (MDS) to visualise the similarity between the different RNA-seq samples (Figure 2). The MDS analyses confirmed the high reproducibility of all biological replicates that cluster closer together within each sample type than between sample types. We also evaluated the reliability of our dataset by visualizing the read coverage of the only two known intron-containing genes in the *T. brucei* genome: Tb927.3.3160 (Nuclear poly(A) polymerase 1) and Tb927.8.1510 (ATP-dependent RNA helicase DBP2B). The visualisations in Figure 3 and Figure 4 show that the intron containing regions of the two genes have a sudden drop with little or no coverage in the polysomal samples (yellow tracks) relative to the total and sub-polysomal samples (blue and purple tracks) in both the BSF and PCF samples.

Comparison with previous work

We compared our results with those of Antwi *et al.*[12] that describes a similar approach to that used in this study. We first analyzed our dataset by counting reads aligned to coding sequence regions (CDS) only. After read normalization in edgeR, we computed the percentages of transcripts bound by the polysome for each gene. These values were then corrected for the relative proportions of mRNA found in polysome fractions (70%) and sub-polysomal fractions (30%) to mimic the analysis pipeline described in 12 as closely as possible. The percentage of transcripts bound by polysomes from our study was then compared with those reported in Table S1 of 12 (Figure 5). The comparison showed a stronger correlation in the PCF life stage (R^2=0.91) than in the BSF life stage (R^2=0.74).

Bias correction

Before further analysing our datasets, we examined GC content bias and length bias in our read counts as those have been reported to affect RNA-seq experiments[42,58,59]. The data in (Figure 6 and Figure 7) show that GC content and length biases affect our dataset in a sample-specific way, especially between the sub-polysomal samples (green) relative to the polysomal (blue) and total (grey) samples. We corrected the read counts for these biases and normalized the read counts using the conditional quantile normalization method implemented in the cqn R package[42].

Differential abundance analysis

Before proceeding to the differential abundance analysis, we visualized the whole dataset with a dimensionality reduction technique. Using an ANOVA-like test implemented in edgeR, we found transcripts that are differentially abundant between any of the groups, without biasing before-hand which groups might be different. We then took the median value of each biological replicate for each gene and applied a radial visualization plot that uses a polar coordinate system to visualize the dataset. Sample types are like hours on the clock-face (i.e. related to the angle of the polar coordinate system) and the orthogonal axis (i.e. the distance from the centre) relates to the relative abundance of a gene across the samples. This analysis showed a strong signature for the BSF and PCF sub-polysomal samples, where many transcripts showed the greatest differential abundance relative to all of the other samples (Figure 8, blue and orange gene dots).

To try to gain insight into this signature, we performed a cluster analysis. We first determined the optimal number of clusters (n=4) with the elbow approach (Figure 9), and then applied a k-means clustering algorithm to divide our dataset into 4 clusters (*Extended data*: Table 3[60]). Cluster 1: gene transcripts that are more abundant in PCF versus BSF samples. Cluster 2: gene transcripts that are more abundant in BSF and PCF sub-polysomal samples than in all other samples. Cluster 3: gene transcripts that are more abundant in BSF versus PCF samples. Cluster 4: gene transcripts that are less abundant in BSF and PCF sub-polysomal samples than in all other samples. This clustering analysis confirmed the presence of a group of genes (Cluster 2, n= 3356) with the highest read counts in the BSF and PCF sub-polysomal samples relative to all other samples (Figure 10).

To assign possible biological functions to the clusters, we performed a GO-term enrichment analysis across the four clusters. We only retained GO terms that were enriched in at most two of the four clusters, and those with false discovery rates of >1%. This analysis, visualized in (Figure 11), showed

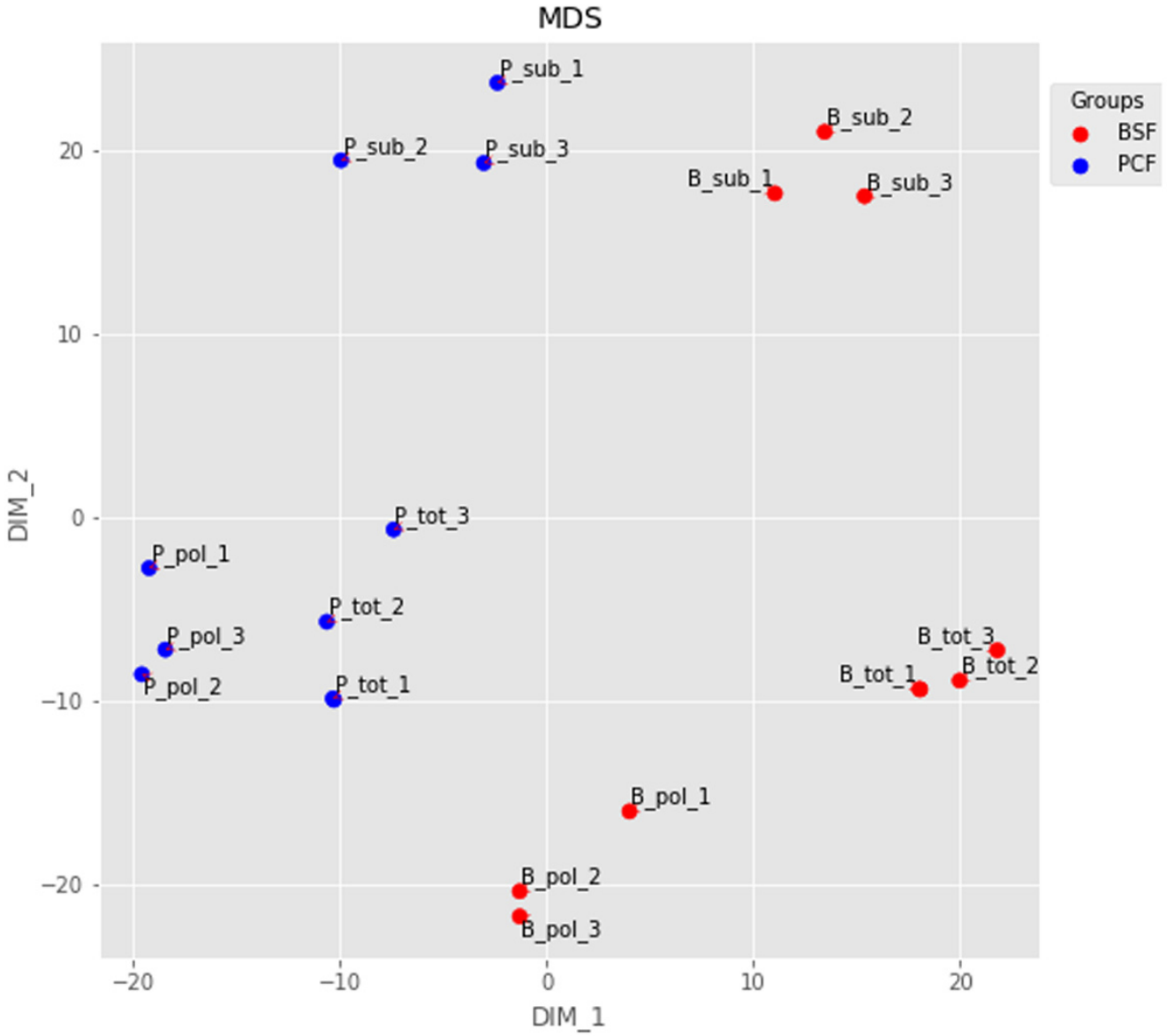

Figure 2. Dimensionality reduction. The output of a multidimensional scaling analysis of the top 500 transcripts for: B_tot_1-3 = Bloodstream Form (BSF) total messenger RNA (mRNA) from samples 1-3; B_sub_1-3 = BSF sub-polysomal mRNA from samples 1-3; B_pol_1-3 = BSF polysomal mRNA from samples 1-3; P_tot_1-3 = Procyclic Form (PCF) total mRNA from samples 1-3; P_sub_1-3 = PCF sub-polysomal mRNA from samples 1-3; P_pol_1-3 = PCF polysomal mRNA from samples 1-3.

that the transcripts in Cluster 2 (C2) are highly enriched for those encoding mRNA binding proteins. Interestingly, the average half-life of the transcripts in Cluster 2 are the shortest in the BSF and the PCF life stages, when compared to the mRNA half-lives of the transcripts in the other clusters (Table 1 and Figure 12). We then asked if any of the clusters are particularly enriched for the long non-coding genes identified in 31 and found they are mostly enriched in Cluster 2 (Table 2). Cluster 2 also has the highest number of two other classes of non-coding mRNAs: the snoRNAs and H/ACA-like snoRNAs (Table 2).

We then focused on the analysis of the (cluster 2) transcripts enriched in the sub-polysomal samples. We created two models to test for differential abundance between the sub-polysomal and polysomal samples in the BSF (*Extended data*: Table 4[60])

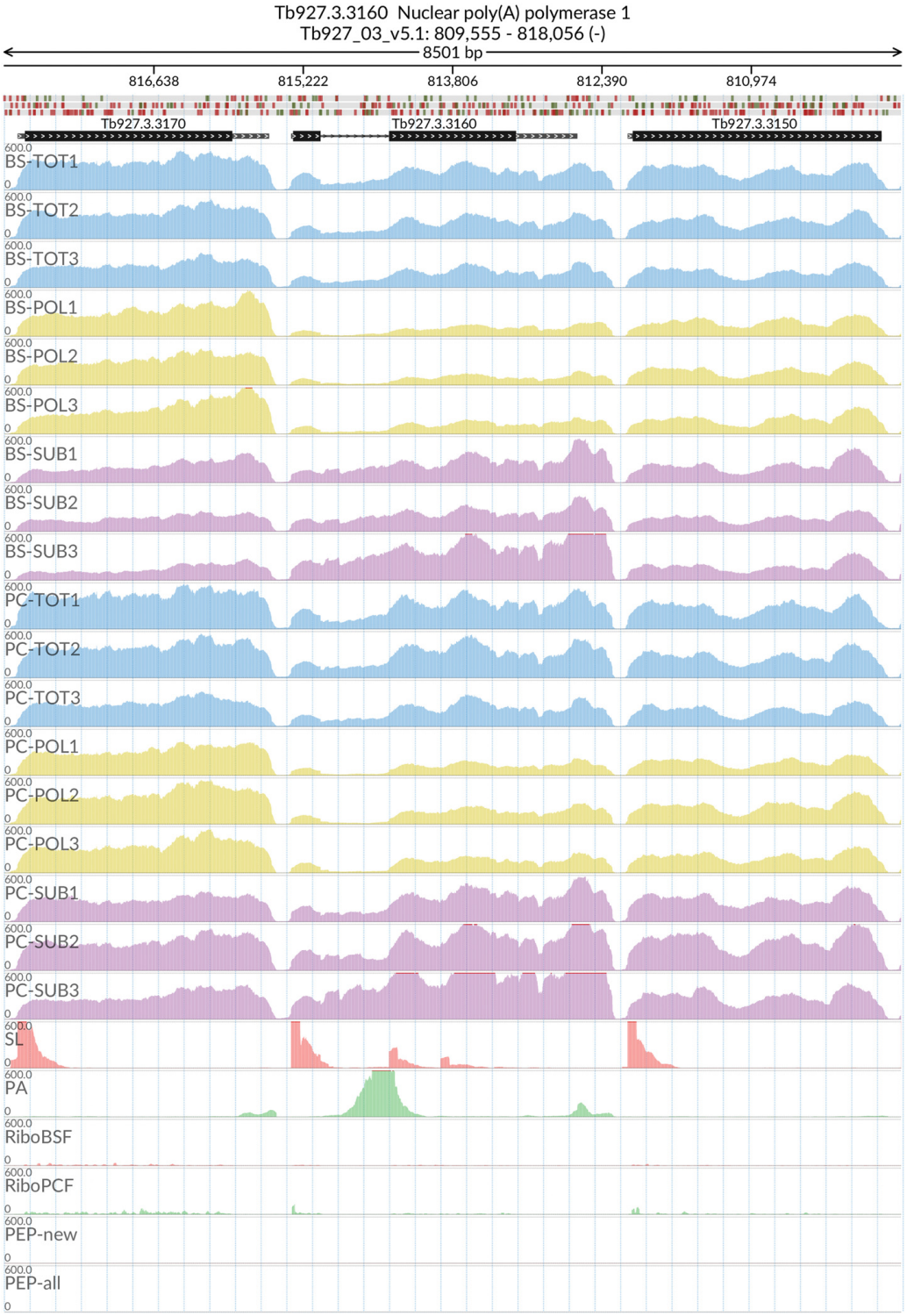

Figure 3. Genome coverage for Tb927.3.3160. For the intron containing gene Tb927.3.3160 (Nuclear poly(A) polymerase 1) the figure shows the genome coverage for the total (TOT), polysomal (POL), and subpolysomal (SUB) samples (biological replicates 1 to 3) of the bloodstream (BS) and procyclic (PC) form life stages. The figure also reports the genome coverage of the Splice Leader (SL) and poly(A) mRNA tails and/or poly(A) genomic tract (PA) containing reads assembled from the samples. Also shown are the ribosome profiling reads for the Bloodstream Form (RiboBSF) and Procyclic Form (RiboPCF) life stages as described in Vasquez *et al*. 2014. The last two genomic tracks report the peptide identifications for new predicted open reading frames (PEP-new) and for all the open reading frames (PEP-all) in TritrypDB. The maximum height of each of the gene tracks is reported on the top left of each track. The top of the figure shows an ideogram of the gene structures. The three grey genomic tracks at the top report ATG codons in green and stop codons in red.

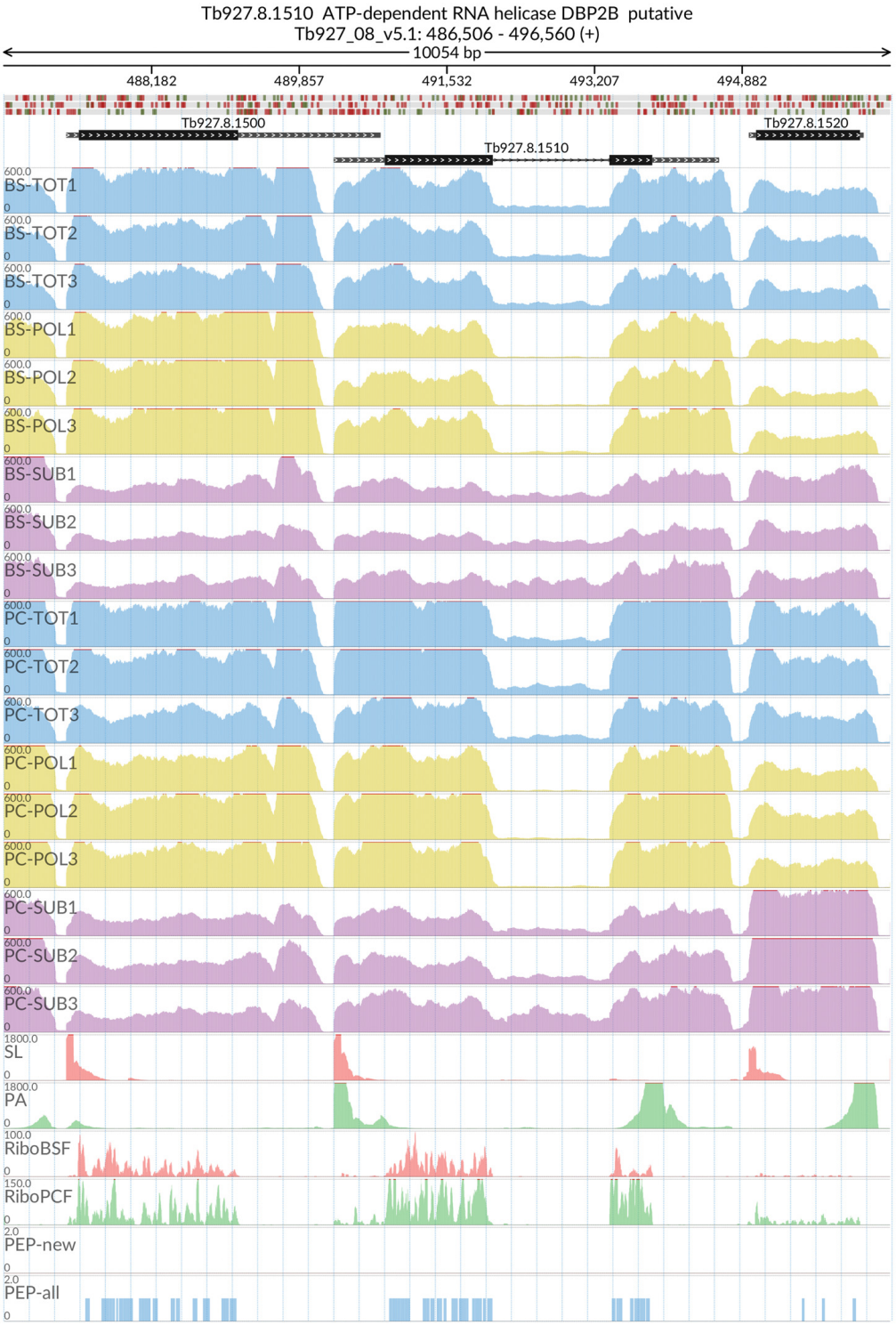

Figure 4. Genome coverage for Tb927.8.1510. For the intron containing gene Tb927.8.1510 (ATP-dependent RNA helicase DBP2B) the figure shows the genome coverage for the total (TOT), polysomal (POL), and subpolysomal (SUB) samples (biological replicates 1 to 3) of the bloodstream (BS) and procyclic (PC) form life stages. The figure also reports the genome coverage of the Splice Leader (SL) and poly(A) mRNA tails and/or poly(A) genomic tract (PA) containing reads assembled from the samples. Also shown are the ribosome profiling reads for the Bloodstream Form (RiboBSF) and Procyclic Form (RiboPCF) life stages as described in Vasquez *et al*. 2014. The last two genomic tracks report the peptide identifications for new predicted open reading frames (PEP-new) and for all the open reading frames (PEP-all) in TritrypDB. The maximum height of each of the gene tracks is reported on the top left of each track. The top of the figure shows an ideogram of the gene structures. The three grey genomic tracks at the top report ATG codons in green and stop codons in red.

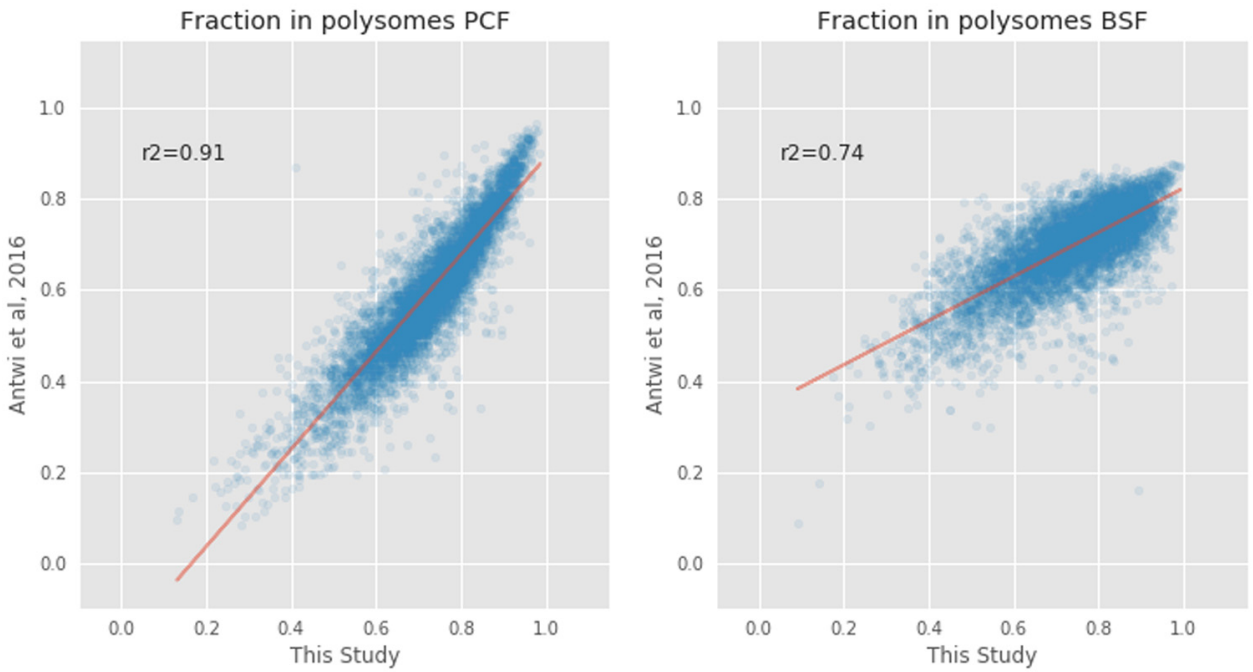

Figure 5. Comparison of the polysomal transcripts between this study and that of Antwi *et al.*[12]. The proportions of messenger RNA (mRNA) transcripts (blue circles) found in polysomal fractions in 12 (y=axis) and in this study (x-axis) in Procyclic From (PCF, left plot) and Blood Stream From (BSF, right plot) samples. The Pearson correlation coefficients (r^2) are 0.91 and 0.74, respectively.

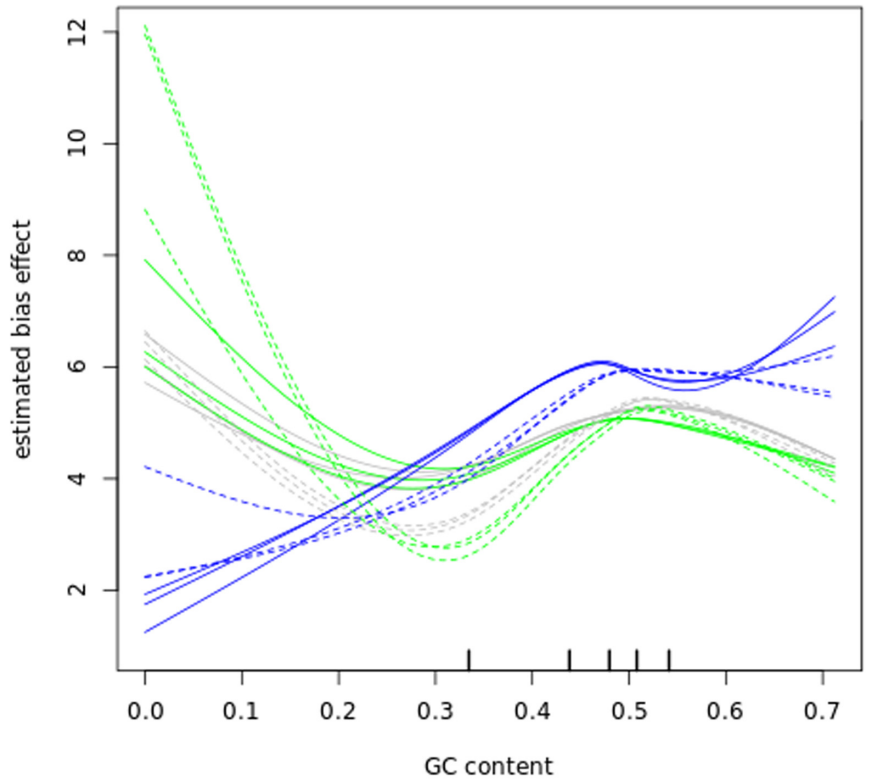

Figure 6. GC bias. A plot of gene transcript guanine-cytosine (GC) content percentage (x-axis) versus the log2 Reads Per Kilobase of fragment, per Million mapped reads (FPKM) estimated bias effect (y-axis) of the bloodstream (B, solid lines) and procyclic (P, dashed lines) samples. The blue lines plot the sub-polysomal (sub) samples, the green lines plot the polysomal (pol) samples and the grey lines plot total (tot) sample bias effects.

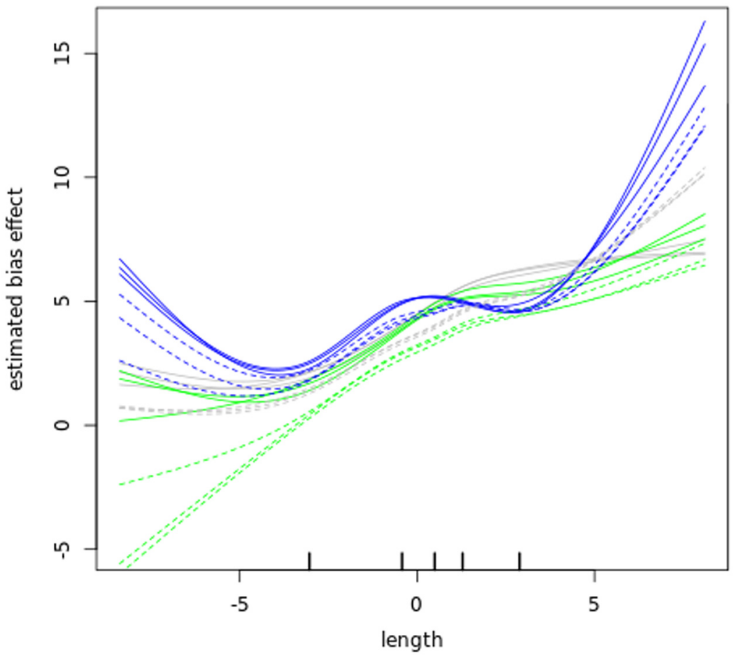

Figure 7. Length bias. A plot of gene transcript length (log2 kilobase) along the x-axis versus the estimated log2 Reads Per Kilobase of fragment, per Million mapped reads (FPKM) bias effect (y-axis) of the bloodstream (B, solid lines) and procyclic (P, dashed lines) samples. The blue lines plot the sub-polysomal (sub) samples, the green lines plot the polysomal (pol) samples and the grey line plots total (tot) sample bias effects.

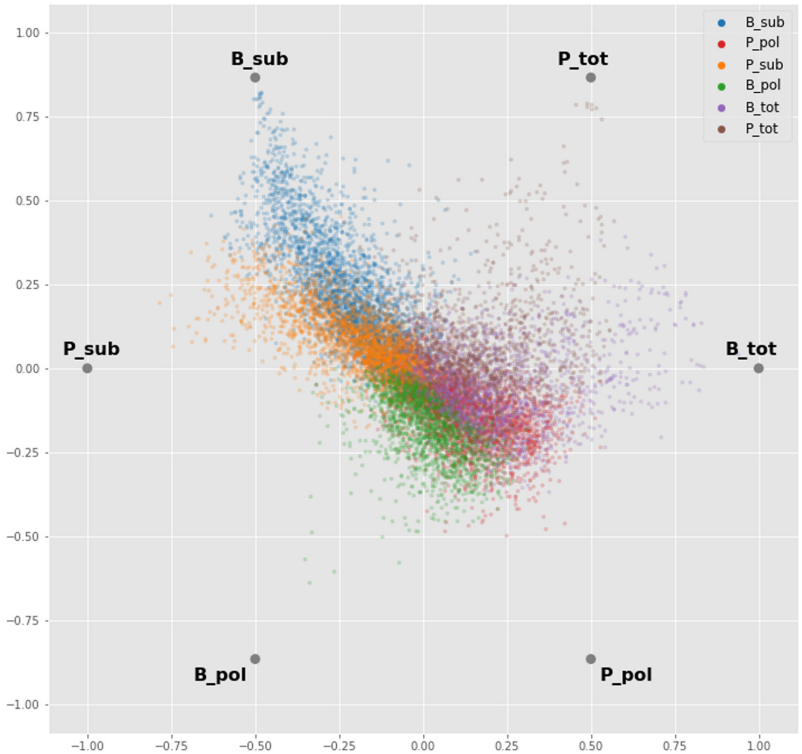

Figure 8. Radial Visualization. A plot from the RadViz algorithm applied to the experimental samples arrayed uniformly around the circumference of a circle. Each gene (dots) is plotted on the interior of the circle such that the distance of the dot on a line from the circumference to the centre is proportional to the gene counts. The dot is colour coded according to the sample where it has the maximum read count value. P = procyclic form, B = bloodstream form, sub = sub-polysomal transcripts, pol = polysomal transcripts, tot = total transcripts.

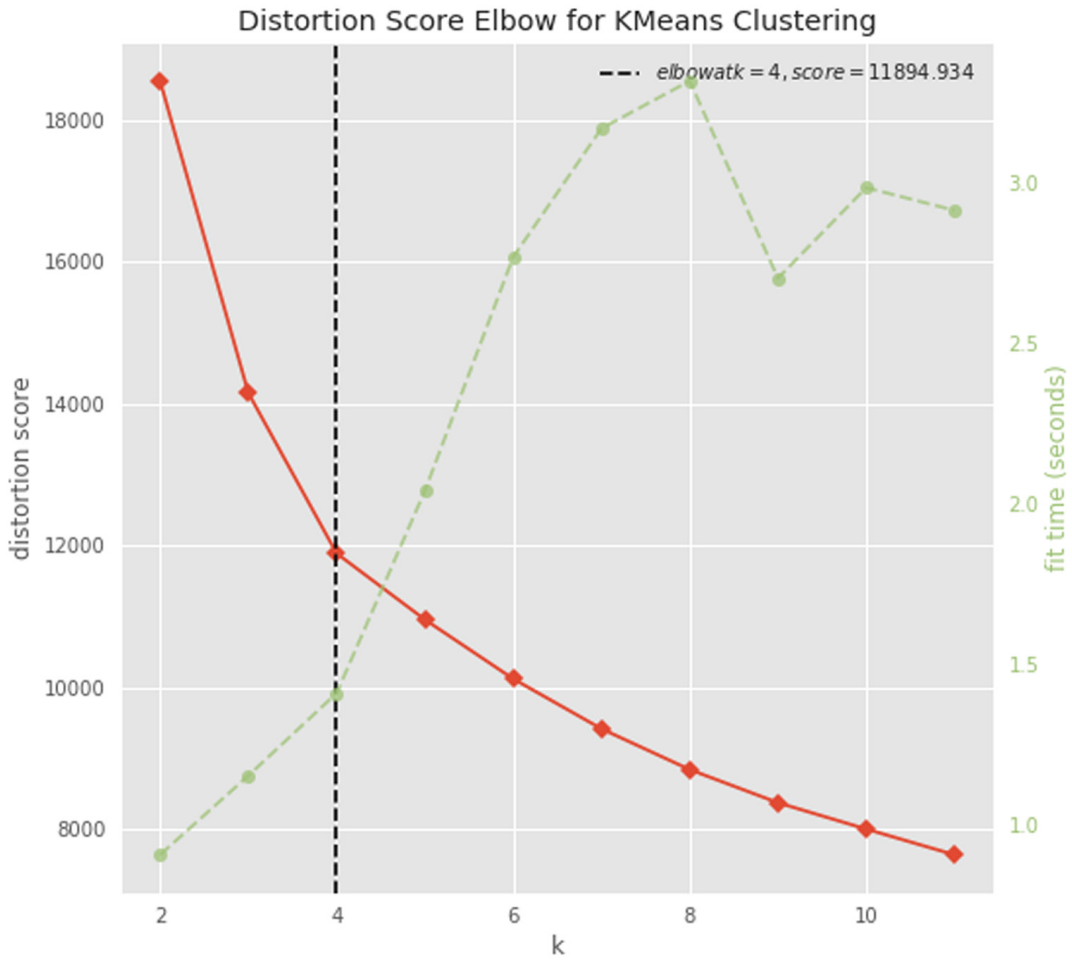

Figure 9. Determining the optimal number of clusters. A plot of the number of clusters tested (K) on the x-axis and the clustering distortion score (the sum of square distances from each point to its assigned cluster center) on the y-axis. The figure also displays the amount of time needed to train the clustering model per K as a dashed green line. If the line chart resembles an arm, then the "elbow" (the point of inflection on the curve) is a good indication that the underlying model fits best at that point (https://www.scikit-yb.org/en/latest/api/cluster/elbow.html).

and PCF (*Extended data*: Table 5[60]) life stages. As illustrated in Figure 13, several long non-coding genes are more abundant in the sub-polysomal samples with respect to the polysomal samples, including the *grumpy* transcript (Figure 14) that sits at the 5' end of RBP7A (Tb927.10.12080) and has been shown to be important for the progression from the slender form to the stumpy form of the parasite[26]. The *grumpy* transcript made us wonder which other sub-polysome enriched transcripts might have a lncRNA at the 5' end and might be associated with this life stage transition. We identified two candidate genes: RBP10 (Tb927.8.2780) with the lncRNA KS17gene_1749a (Figure 15) and REG9.1 (Tb927.11.14220) with the lncRNA KS17gene_4296a (Figure 16), both of which have been previously associated with the transition between the BSF and PCF life stages[61,62].

Expression Analysis of lncRNAs and surrounding genes

We then asked if we could find evidence of coregulation between the lncRNAs and the genes at their 5' or 3'. To this aim, we first mapped again our dataset to the *T. brucei* genome considering only the coding sequence (CDS) of protein coding genes and the lncRNAs. We decided to consider only the CDSs for two reasons. First, several UTRs are not well annotated in T. brucei and, second, multiple lncRNAs overlap with UTR regions. We than created a new model to test for differential abundances between BSF and PCF sub-polysomal samples versus BSF and PCF polysomal samples. Finally, we reported the log fold change of lncRNAs (sub-polysomal vs polysomal samples) along with the log fold changes of the transcripts of genes at their 5' or 3'. This allowed us to use a McNemar's test and observe a statistical significant association between

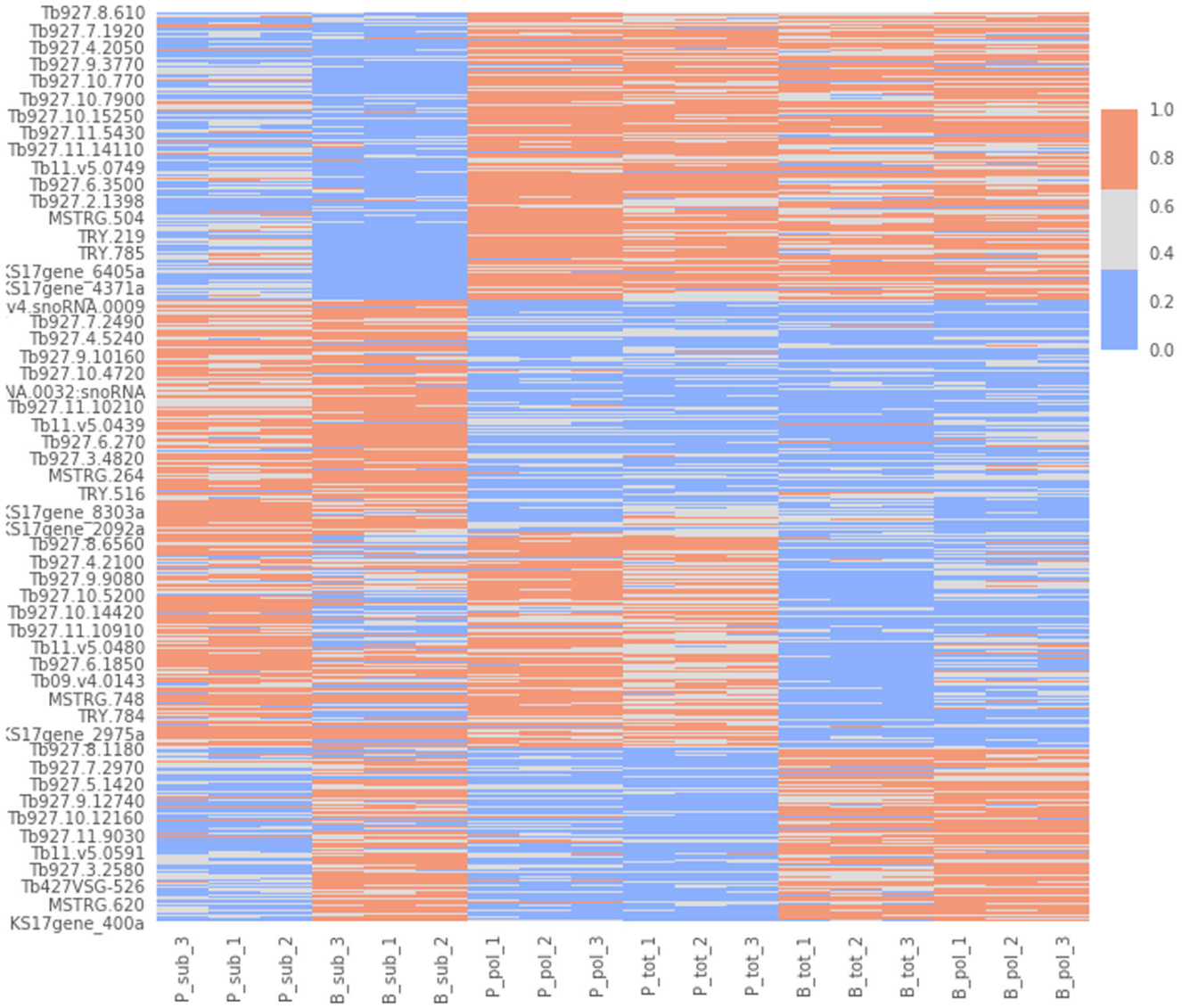

Figure 10. Cluster visualization. A heatmap of the normalized gene count values for the biological replicates (x-axis) against gene identifications (IDs, y-axis). The figure uses three colour codes (colour bar, top right) to visualize the intensity of the normalized read counts (red - highest, gray - middle, blue - lowest). The biological replicates are listed in the format of [B/P]_[tot/pol/sub]_[1/2/3] where B: bloodstream form, P: procyclic form, tot: total RNA sample, pol: polysomal sample, sub: subpolysomal sample, 1,2,3: biological replicate identifiers.

the differential abundance of the lncRNAs and the genes at their 5' (pval $1e^{-16}$). In particular, we observed that lncRNAs that are more abundant in the polysomal fractions are more likely to have a gene at their 5' that is more abundant in the polysomal fractions as well (Figure 17). We could find a similar association between the lncRNAs and the genes at their 3', but several order of magnitude weaker (pval $1e^{-3}$). The GO term analysis of those genes at the 5' of lncRNAs, where both the lncRNAs and the 5' genes are more abundant in the polysomal fractions, showed an enrichment for the following GO terms: posttranscriptional regulation of gene expression; cytoplasm; glycosome and mRNA binding. Since the GO term enrichment analysis highlighted a possible role of lncRNAs in regulating transcripts involved in post-transcriptional regulation of gene expression, we intersected the lncRNAs surrounding genes with a list of 322 potential post-transcriptional regulators in *T. brucei*[63] (Extended data: Table 6[60]).

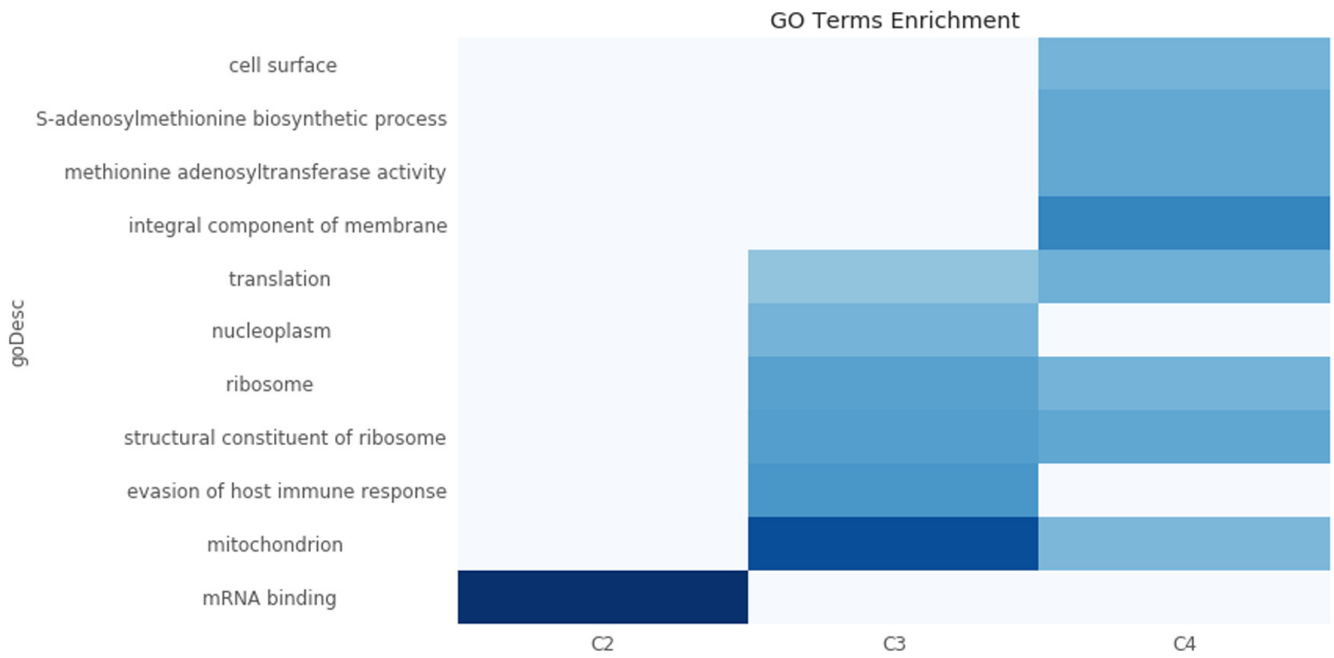

Figure 11. GO term enrichment analysis. A heatmap of the -log10 p-value of the Gene Ontology (GO) term enrichment test. The clusters are plotted in the x-axis and the top enriched GO terms on the y-axis. The -log10 p-value is colour coded according to the colormap on the bottom-right of the plot. The GO terms enriched in >2 clusters have been removed. The cluster C1 (underrepresented in sub-polysomal samples) has been removed from the figure for visualization as it reports the longest list of enriched GO terms (n=41). C2: Cluster 2, genes with the highest gene counts in the Bloodstream From (BSF) and Procyclic Form (PCF) sub-polysomal samples. C3: Cluster 3, genes that are more highly present in BSF samples with respect to PCF samples. C4: Cluster 4, genes with a lower abundance in the sub-polysomal BSF and PCF samples with respect to all the other samples.

Table 1. Half life report. Median half-lives for each cluster of messenger RNA (mRNAs) as extracted from Antwi *et al.*[12].

Cluster	half-life(PCF)	half-life(BSF)
4	27.0	14.3
1	21.0	10.8
2	15.0	10.6
3	17.0	11.6

Identification of new protein coding genes

We were interested in evaluating whether there is proteomic evidence for the new hypothetical protein-coding genes identified in our dataset. To achieve this, we analyzed our protein half-life dataset[48], which provide deep total BSF and PCF proteomes derived from a total of 480 LC-MS/MS runs, running MaxQuant with a database of open reading frames (ORFs) for the TREU927 genome downloaded from TryTripDB. The genomic coordinates of the ORF peptides were then intersected with the genomic coordinates of the hypothetical new protein coding genes. Further, we filtered out unannotated genes in the main 11 chromosomes of *T. brucei* which lacked a splice leader site and/or ribo-seq data. This analysis led to the identification of 11 new hypothetical protein coding genes reported in *Extended data*: Table 7[60].

As examples, two of these hypothetical protein coding genes (TRY.375 and MSTRG.94) are described further.

TRY.375

The start and end of the putative gene were designated at Tb927_07_v5.1:828803.. 830064 by Spliced Leader (SL)/Poly-A (PA) mapping. The putative TRY.375 gene (Figure 18) contains a predicted open reading frame of 522 base pairs encoding for a protein of 173 amino acids (19.51 kDa). The TRY.375 protein product is predicted to have an uncleaved signal peptide and three transmembrane domains. Blastp analysis of the protein product returned low percentage identity (<50%) matches with genes in *T. grayi* (DQ04_00451000), *T. conorhini* (accession: XP_029230363.1) and *T. theileri* (TM35_000192250). Synteny analysis of the TRY.375 locus performed at TryTripDB revealed another gene (TevSTIB805.7.3380) in the *T. evansi* genome with 100% identity with the predicted TRY.375 gene product. Also, a tblastn search of the *TRY.375* predicted gene identified 2 more hits with 100% identity in the genomes of *T. brucei* 427_2018, 427 (Tb427) and *T. brucei gambiense* DAL972 (Tbg972), corresponding to unannotated regions in these genomes. We propose that TRY.375

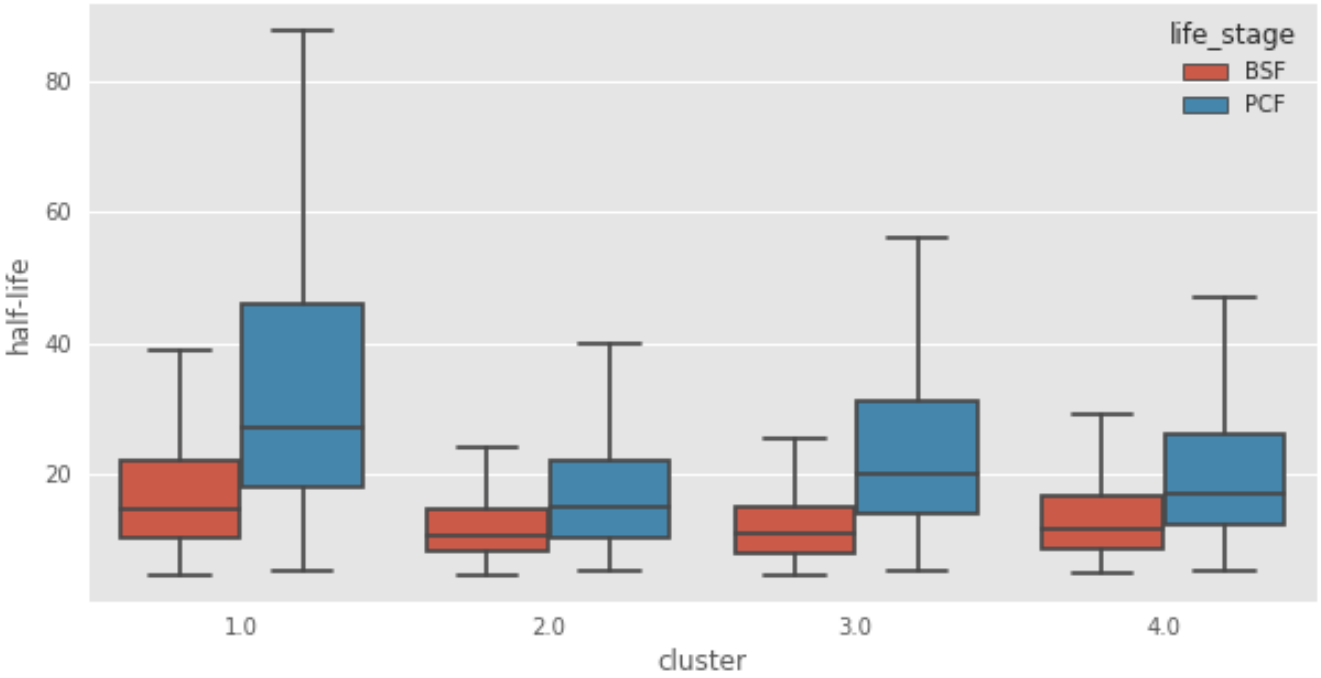

Figure 12. Transcript half-life. Boxplots of messenger RNA (mRNA) half-life in minutes (y-axis) for the genes assigned to the clusters reported in the x-axis for the Bloodstream From (BSF, red) and Procyclic Form (PCF, blue) life stages. 1: Cluster 1, transcripts underrepresented in BSF and PCF sub-polysomal samples, 2: Cluster 2, genes with the highest read counts in the BSF and PCF sub-polysomal samples 3: Cluster 3, genes that are more highly present in BSF samples with respect to PCF samples; 4: Cluster 4, genes with a lower abundance in the sub-polysomal BSF and PCF samples with respect to all the other samples.

Table 2. Non coding mRNA counts. The number of Small nucleolar RNAs (snoRNAs), H/ACA-like containing box snoRNAs (H/ACA-like snoRNAs) and long non-coding RNAs (lncRNAs) identified in each cluster.

Cluster	snoRNAs	H/ACA-like snoRNAs	lncRNAs
1	17	5	405
2	180	43	473
3	66	12	330
4	20	6	206

is a novel transmembrane-protein coding gene present in *T. brucei* and *T. evansi.*

MSTRG.94

Peptides corresponding to potential new gene MSTRG.94 (Figure 19) mapped with high confidence to 6 regions within the span Tb927_02_v5.1:592500..617500. Investigation of this section of chromosome 2 revealed it is highly repetitive and contains 6 copies of a 65kDa Invariant Surface Glycoprotein (ISG65) gene with a pairwise protein Identities computed by Clustal Omega between 73% and 99%. This suggests that what had previously been assumed to be untranslated intergenic regions of DNA may in fact encode for protein. SL and PA mapping allowed us to define 6 MSTRG.94 gene boundaries as described in *Extended data*: Table 7[60]. All of these putative gene regions were identical and we have designated them MSTRG.94_1 through MSTRG.94_6. The putative MSTRG.94 genes contain a predicted ORF of 378 base pairs encoding for a protein of 125 amino acids (14.17 kDa). The predicted protein does not contain any transmembrane domains or signal peptides. A tblastn search with the ORF sequence against trypanosome genomes revealed matching sequences in the genomes of Tb427 and *T. evansi*. As with Tb927, the sequences appear between copies of the ISG65 genes in chromosome 2. In Tb427 the sequences are annotated as hypothetical proteins and in *T. evansi* as unspecified products, while in Tbg972 the regions are unannotated. The transcript seems to be preferentially expressed in BSF form (Figure 19).

Discussion

In this paper we present RNA-seq data on the total, polysomal and sub-polysomal mRNA contents of *T. brucei* bloodstream and procyclic form life stages. Comparison with similar experiments performed earlier by Antwi *et al.*[12] showed good experimental reproducibility between the PCF life stage data (r^2=0.9) and BSF life stage data (r^2=0.7) (Figure 5). A possible source of discrepancy may be different cell culture protocols for the BSF cells. Nevertheless, our datasets showed very

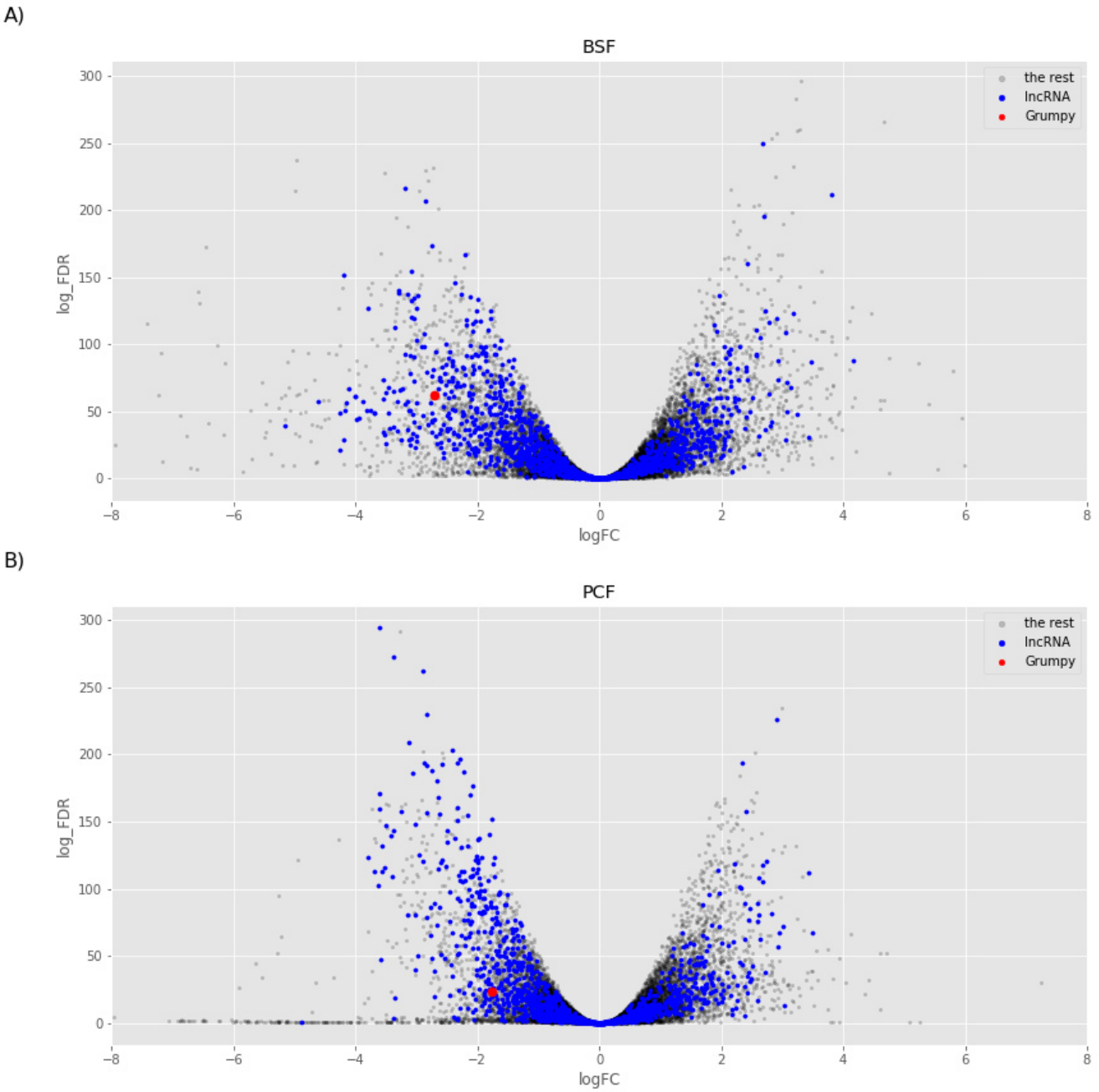

Figure 13. Sub-polysome abundance test. The volcano plots report the log2 fold change (logFC) on the x-axis and the minus log10 of the false discovery rate (log_FDR) on the y-axis obtained from the comparison of the sub-polysomal samples with the polysomal samples for Bloodstream Form (BSF, **A**) and Procyclic Form (PCF, **B**) samples. Blue dots highlight the long non-coding RNAs (lncRNA), red dot highlights the *grumpy* gene described Guegan *et al.*[31], and grey dots highlight the rest of the genes in the sample.

good reproducibility (Figure 2), and we were successful in identifying a pool of efficiently transcribed and spliced mRNAs. This is demonstrated by the virtual absence in the polysomal fractions of reads covering the intron regions of the two experimentally validated intron containing genes (Figure 3 and Figure 4)[64].

By using clustering and dimensionality reduction techniques (Figure 8 and Figure 10), we were able to identify the sub-polysome samples as the most diverse in our dataset. In particular, we found the presence of several long non-coding mRNAs in the sub-polysomal fractions of both BSF and PCF samples (*Extended data*: Table 3[60]). However, some lncRNAs

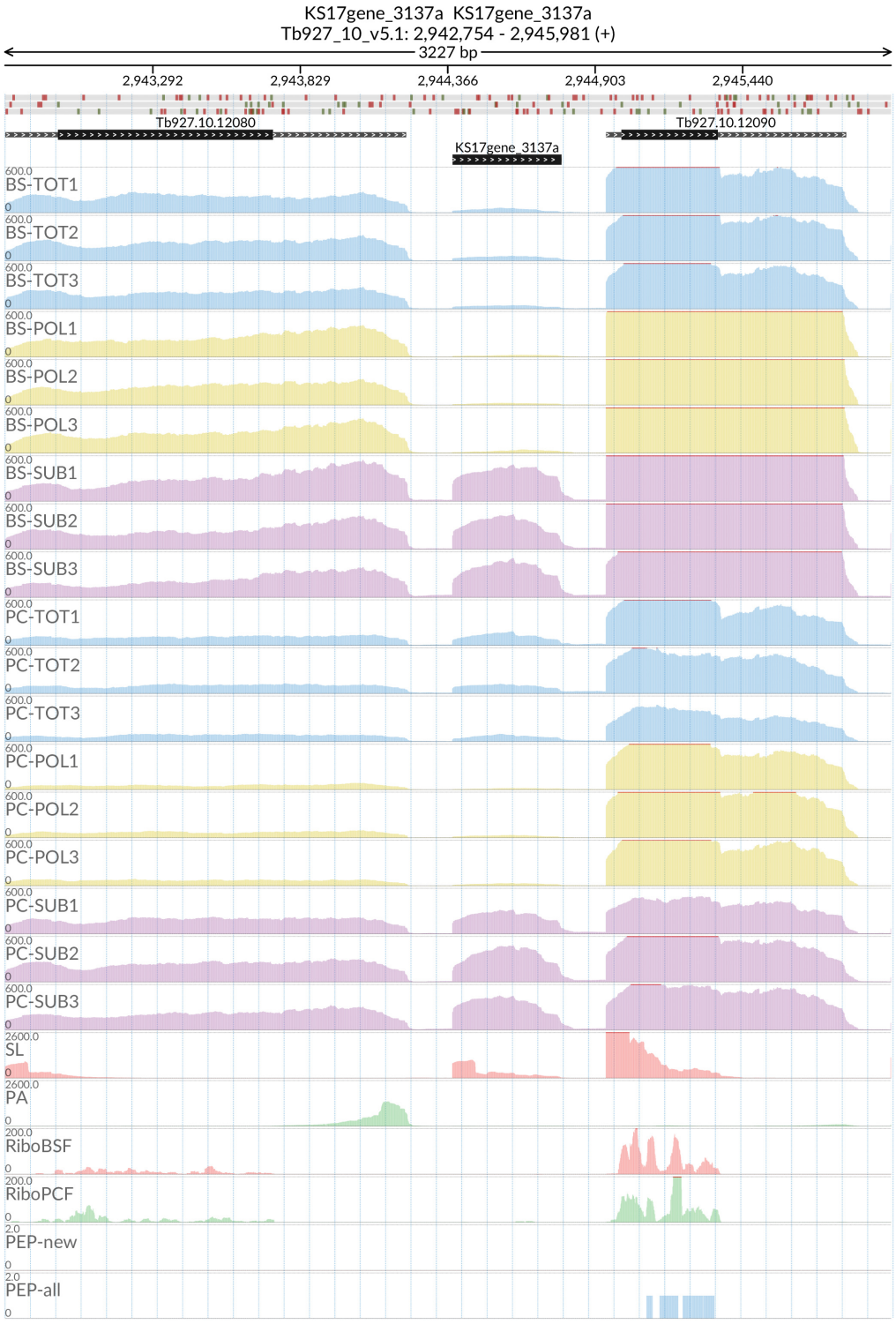

Figure 14. Genome coverage for the *grumpy* gene KS17gene_3137a in the total (TOT), polysomal (POL), and subpolysomal (SUB) samples (biological replicates 1 to 3) of the bloodstream (BS) and procyclic (PC) form life stages. The figure also reports the genome coverage of the Splice Leader (SL) and poly(A) mRNA tails and/or poly(A) genomic tract (PA) containing reads assembled from the samples. Also shown are the ribosome profiling reads for the Bloodstream Form (RiboBSF) and Procyclic Form (RiboPCF) life stages as described in Vasquez *et al.* 2014. The last two genomic tracks report the peptide identifications for new predicted open reading frames (PEP-new) and for all the open reading frames (PEP-all) in TritrypDB. The maximum height of each of the gene tracks is reported on the top left of each track. The top of the figure shows an ideogram of the gene structures. The three grey genomic tracks at the top report ATG codons in green and stop codons in red.

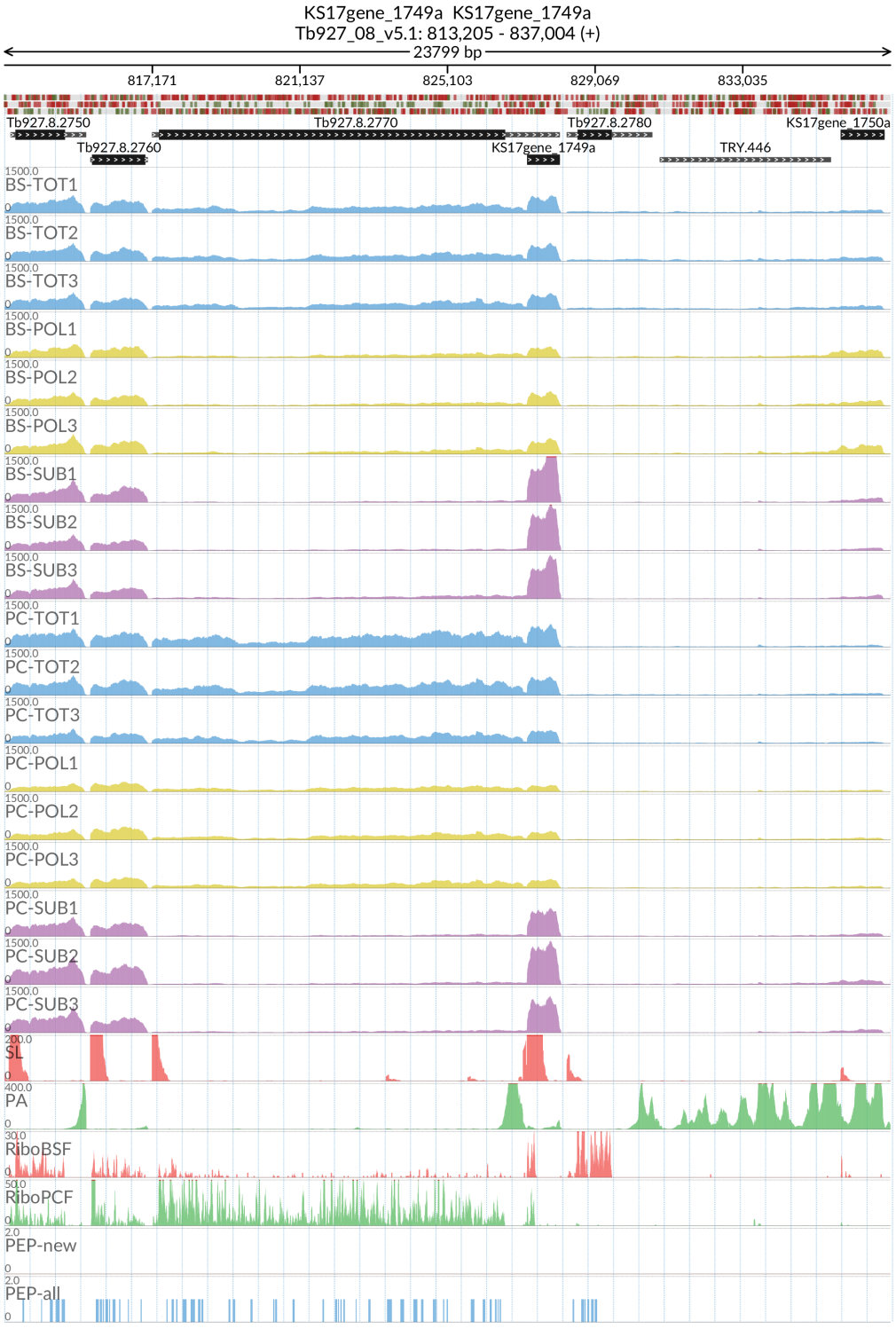

Figure 15. Genome coverage for the long non coding RNA KS17gene_1749a at the 5' of the Tb927.8.2780 (RNA-binding protein RBP10) gene in the total (TOT), polysomal (POL), and subpolysomal (SUB) samples (biological replicates 1 to 3) of the bloodstream (BS) and procyclic (PC) form life stages. The figure also reports the genome coverage of the Splice Leader (SL) and poly(A) mRNA tails and/or poly(A) genomic tract (PA) containing reads assembled from the samples. Also shown are the ribosome profiling reads for the Bloodstream Form (RiboBSF) and Procyclic Form (RiboPCF) life stages as described in Vasquez *et al.* 2014. The last two genomic tracks report the peptide identifications for new predicted open reading frames (PEP-new) and for all the open reading frames (PEP-all) in TritrypDB. The maximum height of each of the gene tracks is reported on the top left of each track. The top of the figure shows an ideogram of the gene structures. The three grey genomic tracks at the top report ATG codons in green and stop codons in red.

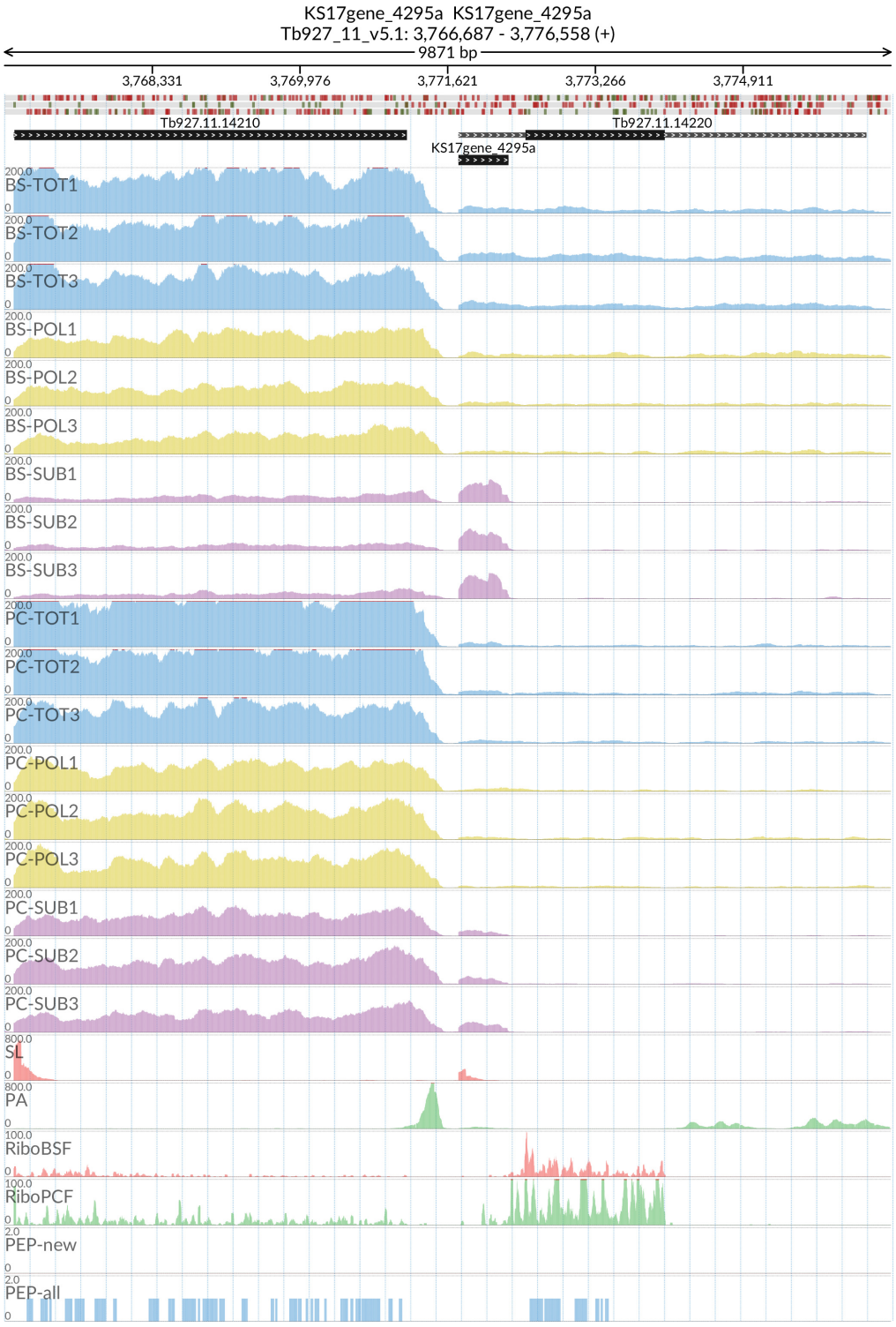

Figure 16. Genome coverage for the long non coding RNA KS17gene_4295a at the 5' of the Tb927.11.14220 (REG9.1) gene in the total (TOT), polysomal (POL), and subpolysomal (SUB) samples (biological replicates 1 to 3) of the bloodstream (BS) and procyclic (PC) form life stages. The figure also reports the genome coverage of the Splice Leader (SL) and poly(A) mRNA tails and/or poly(A) genomic tract (PA) containing reads assembled from the samples. Also shown are the ribosome profiling reads for the Bloodstream Form (RiboBSF) and Procyclic Form (RiboPCF) life stages as described in Vasquez *et al.* 2014. The last two genomic tracks report the peptide identifications for new predicted open reading frames (PEP-new) and for all the open reading frames (PEP-all) in TritrypDB. The maximum height of each of the gene tracks is reported on the top left of each track. The top of the figure shows an ideogram of the gene structures. The three grey genomic tracks at the top report ATG codons in green and stop codons in red.

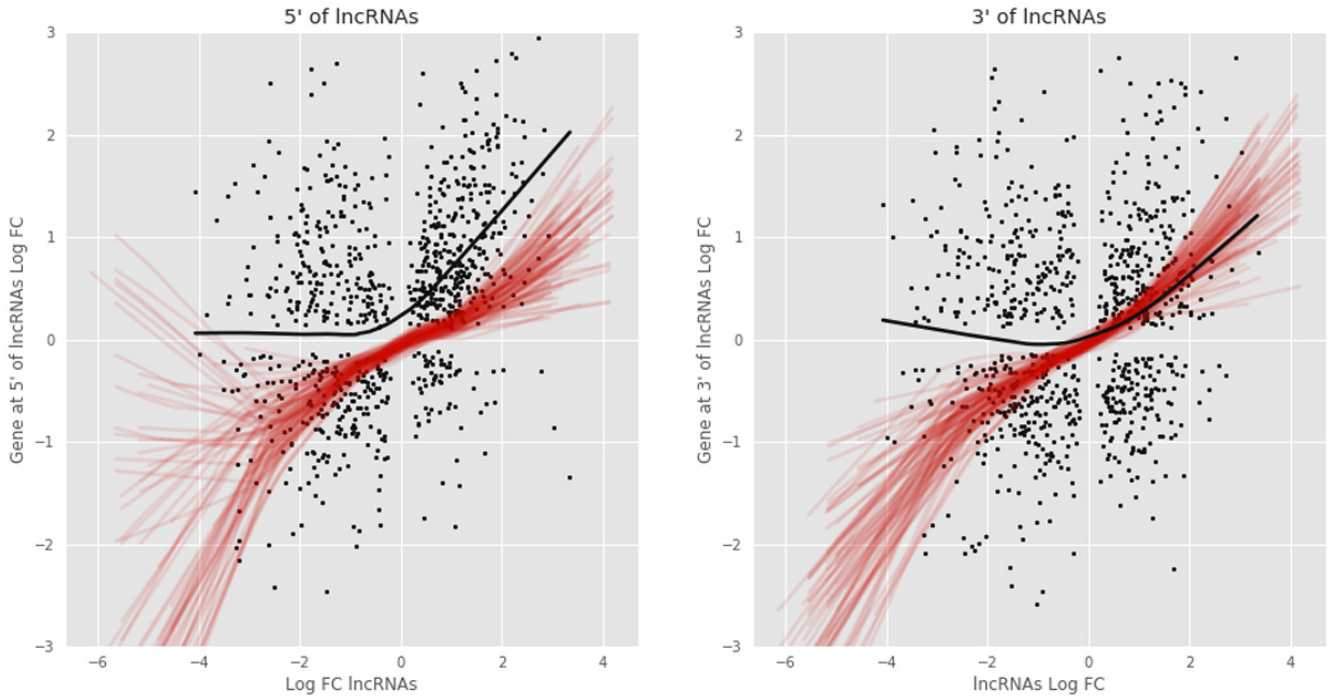

Figure 17. Effect of lncRNAs on the surrounding genes. A) The log2 fold change of the lncRNAs (x axis) between subpolysomal and polysomal fractions (black dots) is plotted against the the log2 fold change of the genes at the 5′ of the lncRNAs (y axis). The black line shows the data trend by fitting a LOWESS regression model. The red lines plot the LOWESS regression models for a selection of random genes (n=1000) and the genes at their 5′. The random gene selection is repeated 100 times. **B**) The log2 fold change of the lncRNAs (x axis) between subpolysomal and polysomal fractions (black dots) is plotted against the the log2 fold change of the genes at the 3′ of the lncRNAs (y axis). The black line shows the data trend by fitting a LOWESS regression model. The red lines plot the LOWESS regression models for a selection of random genes (n=1000) and the genes at their 3′. The random gene selection is repeated 100 times

were also found enriched in the polysomal fractions as already identified in human cells[65]. This class of mRNA has been overlooked in *T. brucei* until recently, and one particular long non-coding mRNA (*grumpy*) has been shown to regulate the transformation from the slender to the stumpy life stage of the parasite[31]. Interestingly, the RNA-binding protein RBP10 (Tb927.8.2780), that has been shown to bind mRNAs and promote their degradation, acts as a molecular switch whereby RBP10 expression in BSF causes differentiation to PCF, while the overexpression in PCF causes differentiation to BSF[61]. While RBP10 itself was not found in our sub-polysome enriched transcript list, the lncRNA (KS17gene_1749a) which is predicted to be at the 5' end of RBP10 may have a similar regulatory function as the *grumpy* lncRNA transcript. Intrigued by these findings, we have assembled a list of lncRNAs along with their surrounding genes at their 5' and 3' ends and reported their differential expression values between the polysomal and sub-polysomal samples (*Extended data*: Table 6[60]). The analysis of these data highlighted a possible co-regulation between the lncRNAs and the genes at their 5' ends, such that when a lncRNA is more abundant in the polysome fractions relative to the sub-polysomal fractions, the same is true of the gene at its 5' end (Figure 17). It is possible that the lncRNAs might influence the transcription efficiency of the proximal genes at their 5' ends, as observed in other organisms[66]. It is possible that lncRNAs might bind to the gene transcript at its 5' end to stabilize it or promote transcription[66]. It may be that these lncRNAs, more abundant in the polysomal fraction of BSF and PCF, regulate genes that are important for the maintenance of such life stages, while the lncRNAs that regulate life stage transitions (like the *grumpy* gene) are targeted to the subpolysomal fractions, possibly for degradation. In any case, we anticipate that the study of lncRNAs transcripts that show differential abundance between the sub-polysomal and polysomal fractions may uncover new mechanisms of transcript stability and regulation in *T. brucei*.

Another class of RNA we found to be enriched in the sub-polysomal fractions are snoRNAs. However some snoRNAs were also detected in the polysomal fraction. The presence of this class of RNA in the polysomal fraction could be explained by contamination, but also by a degradation mechanism. For example, snoRNAs guide the peculiar trypanosome rRNA maturation events, facilitating the methylation and pseudouridylation modification of rRNA[67,68]. Because polyadenylation by snoRNA is a way of marking the RNA for degradation in yeast and humans[69,70], it is possible that a similar mechanism acts in *T. brucei,* and that our poly-A enrichment step has captured this class of RNAs before they have been targeted to the exosome for degradation[71].

Finally, we hope that our dataset will be useful for the annotation of the *T. brucei* genome. Our approach to discover new transcripts in *T. brucei* detected several new transcribed loci.

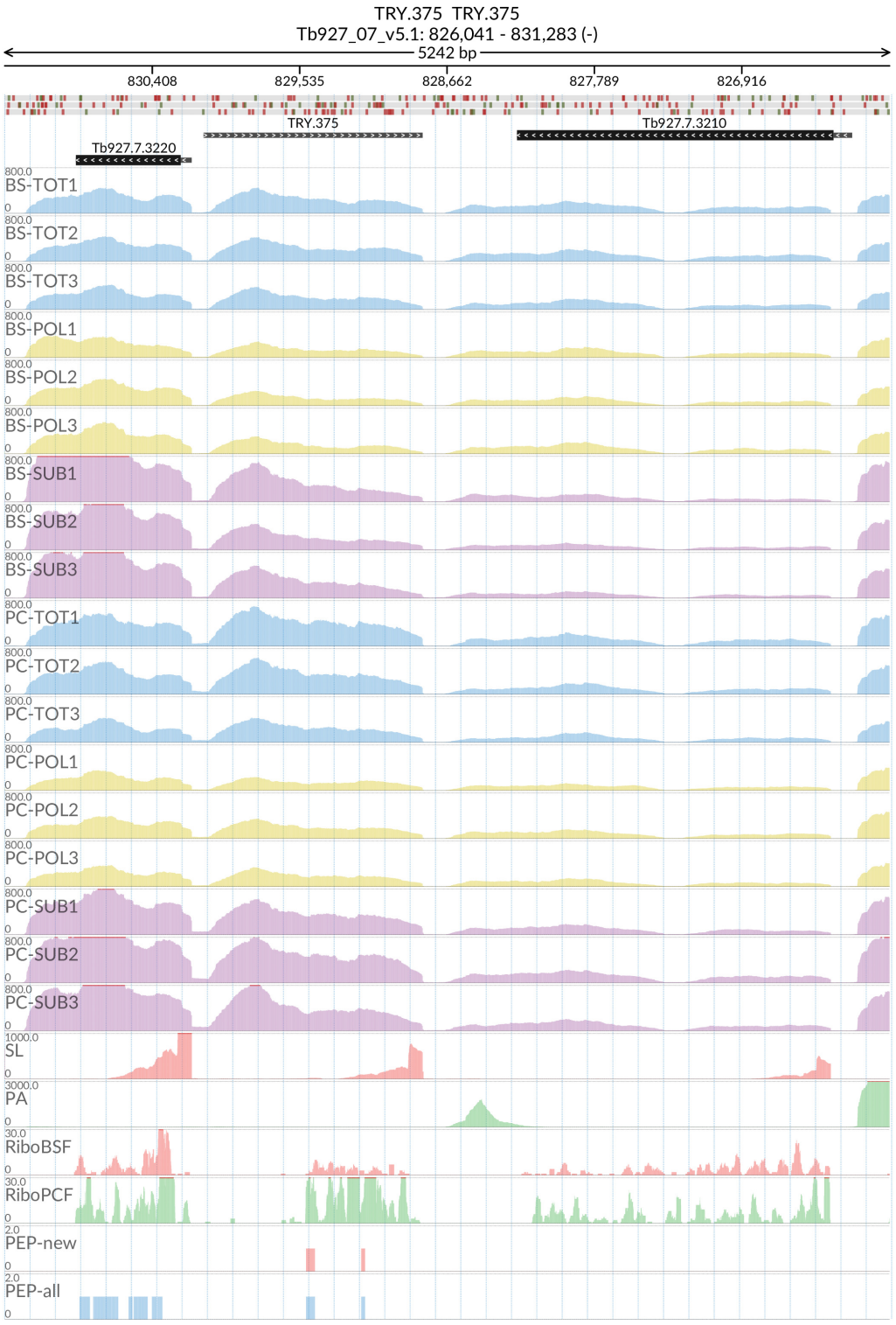

Figure 18. For the new predicted protein coding gene TRY.375, the figure shows the genome coverage for the total (TOT), polysomal (POL), and subpolysomal (SUB) samples (biological replicates 1 to 3) of the bloodstream (BS) and procyclic (PC) form life stages. The figure also reports the genome coverage of the Splice Leader (SL) and poly(A) mRNA tails and/or poly(A) genomic tract (PA) containing reads assembled from the samples. Also shown are the ribosome profiling reads for the Bloodstream Form (RiboBSF) and Procyclic Form (RiboPCF) life stages as described in Vasquez *et al*. 2014. The last two genomic tracks report the peptide identifications for new predicted open reading frames (PEP-new) and for all the open reading frames (PEP-all) in TritrypDB. The maximum height of each of the gene tracks is reported on the top left of each track. The top of the figure shows an ideogram of the gene structures. The three grey genomic tracks at the top report ATG codons in green and stop codons in red.

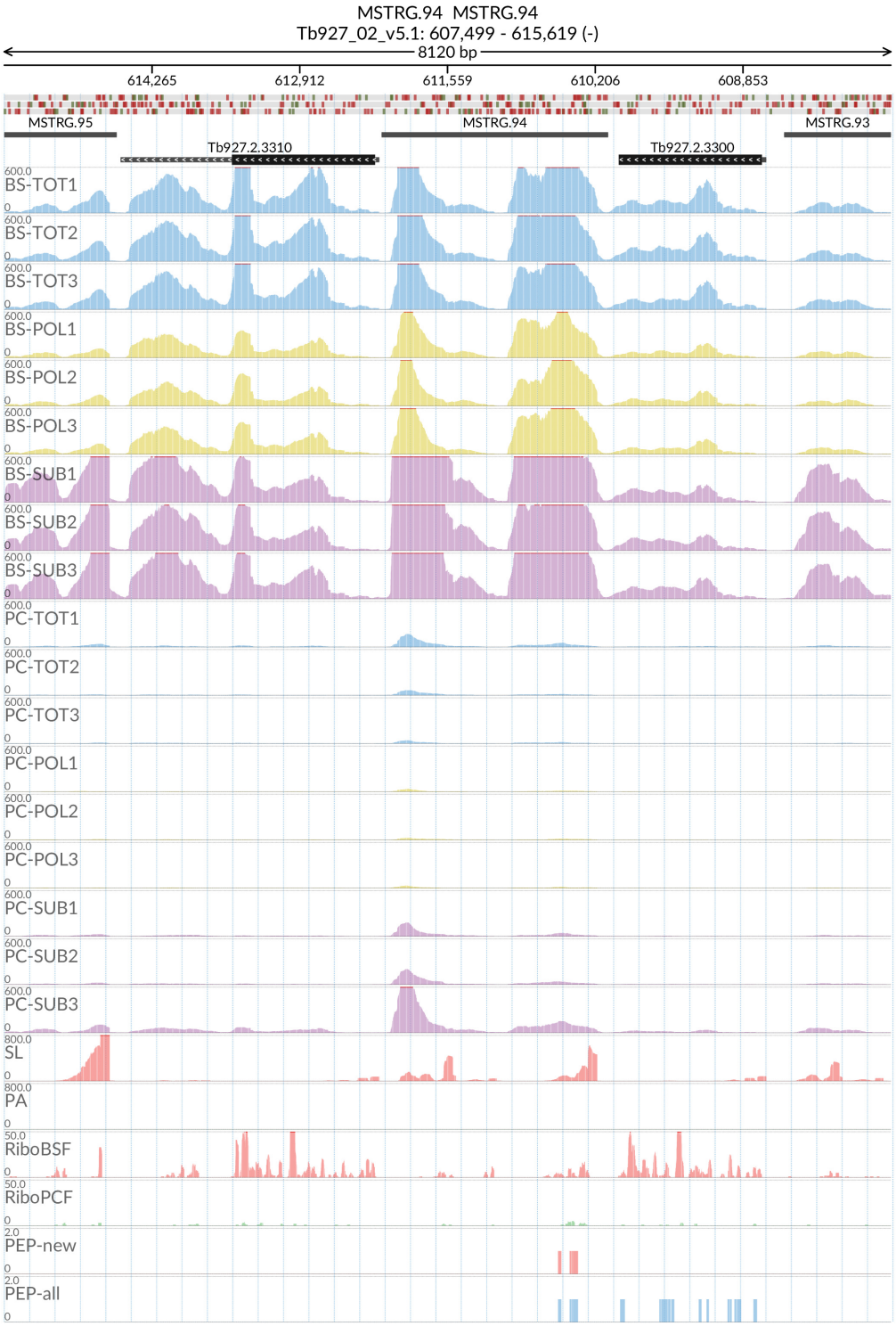

Figure 19. Genome coverage for the new predicted protein coding gene MSTRG.94 in the total (TOT), polysomal (POL), and subpolysomal (SUB) samples (biological replicates 1 to 3) of the bloodstream (BS) and procyclic (PC) form life stages. The figure also reports the genome coverage of the Splice Leader (SL) and poly(A) mRNA tails and/or poly(A) genomic tract (PA) containing reads assembled from the samples. Also shown are the ribosome profiling reads for the Bloodstream Form (RiboBSF) and Procyclic Form (RiboPCF) life stages as described in Vasquez *et al*. 2014. The last two genomic tracks report the peptide identifications for new predicted open reading frames (PEP-new) and for all the open reading frames (PEP-all) in TritrypDB. The maximum height of each of the gene tracks is reported on the top left of each track. The top of the figure shows an ideogram of the gene structures. The three grey genomic tracks at the top report ATG codons in green and stop codons in red.

While most of the transcribed loci represent miss-annotation of putative gene transcripts, we used an unbiased proteomic approach to detect at least 30 new hypothetical protein-coding genes, two of which were further manually annotated here (Figure 18 and Figure 19)

Data availability

Underlying data

All FASTQ files data are deposited at the NCBI SRA database[72] under the bioproject accession number PRJNA634997

Analysis pipeline, links to the raw data and code used to generate the paper figures are available at https://github.com/mtinti/polysome, reproducible using the mybinder badge in GitHub and archived in Zenodo.

Zenodo: mtinti/polysome: Fix Table 6. http://doi.org/10.5281/zenodo.4235160[73]

This project contains the following data:

- (B,P)_(pol, sub, tot)_(1,2,3)
  - counts.txt (The read counts for the genes)
  - counts_CDS.txt (The read counts for the gene coding sequences)
- Figures (The folder containing the figures of the paper)
- Figures_Paper_def.ipynb (The jupyter notebook producing the figures of the paper)
- InData
  - GC_content_927.txt (list of guanine-cytosine content values of the genes in T. brucei)
  - GS_gene_list.txt (list of the hypothetical long non-coding mRNAs in T. brucei according to Guegan F. *et al.* 2020)
  - PTR.txt (list of the genes with a predicted gene expression regulation effect in T. brucei according to Erben, E.D., *et al.* 2014)
  - PolisomeLiterature
    - BSF.csv (The supplementary Table 5 of Antwi *et al.* 2016 for the bloodstream life stage)
    - GeneByLocusTag_Summary.txt (A mapping dictionary to update the gene ids in the supplementary Table 5 of Antwi *et al.* 2016)
    - PCF.csv (The supplementary Table 5 of Antwi *et al.* 2016 for the procyclic life stage)
  - Proteomics
    - peptides_bsf_trim.zip (peptide identification output of MaxQuant in the bloodstream life stage)
    - peptides_pcf_trim.zip (peptide identification output of MaxQuant in the procyclic life stage)
  - TriTrypDB-46_TbruceiTREU927.gff (generic feature format file downloaded from TriTrypDB)
  - TriTrypDB-46_TbruceiTREU927_GO.gaf (Gene Ontology file downloaded from TriTrypDB)
  - TriTrypDB-46_TbruceiTREU927_GO2.gaf (Gene Ontology file modified and used as input for GOATOOLS)
  - go-basic.obo (Ontology file downloaded from http://geneontology.org/docs/download-ontology/)
  - goterm_enrich.txt (list of enriched GO terms in the gene clusters)
  - mRNA_Half_Life
    - mRNAhl_lookup.txt (A mapping dictionary to update the gene ids in the supplementary Table 5 of Antwi *et al.* 2016)
    - mrnaBSFhl.txt (list mRNAs half-lives for bloodstream form as reported in supplementary Table 5 of Antwi *et al.* 2016)
    - mrnaPCFhl.txt (list mRNAs half-lives for procyclic form as reported in supplementary Table 5 of Antwi *et al.* 2016)
  - ribo_counts_927.txt (Read counts for the re-analysis of the ribo-seq dataset)
- Tables (The folder containing the tables of the paper)
- environment.yml (The conda environment file that lists the packages to reproduce the analysis on mybinder)
- make_pipline2.py (python script to assemble the rna-seq analysis pipeline)
- multiQC.ipynb (The jupyter notebook that runs the quality control)
- multiqc_config.yaml (The multiQC configuration file)
- multiqc_fastqc.yaml (The multiQC configuration file for the fastqc package)
- mylib
  - extract_barcodes_def2.py (The python script to extract the RNA-seq reads containing the splice leader sequences or the poly-A tracts)
- polysome_mqc (folder containg the multiQC output files)
- package_versions.txt (a text file listing all the versions of the software used for the analysis)

- postBuild (configuration files for mybinder)
- tb927_3_ks_st_sc_st_tr.gtf (Gene Transfer annotation file of T. brucei listing the new transcribed regions identified in this work)
- tb927_5.fa (Genomic sequences of T. brucei downloaded from TriTrypDB)
- tb927_5.fa.fai (index file Genomic sequences of T. brucei)
- tb927_5.gtf (Gene Transfer annotation file of T. brucei downloaded from TriTrypDB)
- templates
  - scallop.sh (the bash script to run scallop for the identification of new transcribed regions))
  - template_rnaseq.sh (the bash script to run the RNA-seq analysis pipeline)
  - trinity_template.sh (the bash script to run trinity for the identification of new transcribed regions)
- README.md (the github readme file)
- utilities.py (python script with helper functions for the data analysis)
- vars5.txt (list the input parameters for the make_pipline2.py file)
- wcar.png (Wellcome Centre for Anti-Infectives Research logo)

The code and the data used to generate the paper figures that visualise the RNA-seq coverage are available at https://github.com/mtinti/polysome_coverage, https://github.com/mtinti/polysome, reproducible using the mybinder badge in github and archived in zenodo.

Zenodo: mtinti/polysome_coverage: pre-submission. http://doi.org/10.5281/zenodo.4428343[74]

This project contains the following data:

- (B/P)_(pol/sub/tot)_(1/2/3)_sorted_pc_bg.bed (bed graph file for the coverage of the RNA-seq samples)
- Figures_Paper_Coverage.ipynb (The jupyter notebook that produce the coverage images)
- README.md (the GitHub readme file)
- Tb927.8.1510_paper_figures.png (coverage image for the Tb927.8.1510 gene)
- all_927_F_plus_R_SL.bed (bed graph file format for the coverage of the reads containing the spliced-leader sequences)
- all_F_plus_R_PoliA.bed (bed graph file format for the coverage of the reads containing the poli-A tract)
- all_pepe.bed (bed graph file format for the coverage of the peptides identified with mass spectrometry)
- environment.yml (The conda environment file that lists the packages to reproduce the coverage analysis on mybinder)
- new_genes.bed (bed graph file format for the coverage of the peptides identified with mass spectrometry for new predicted protein coding gene)
- package_versions.txt (a text file listing all the versions of the software used for the analysis)
- riboBSF_927.bed (bed graph file format for the coverage of ribo-seq samples in the bloodstream samples)
- riboPCF_927.bed (bed graph file format for the coverage of ribo-seq samples in the procyclic sample)
- svist4getConf (configuration folder for the svist4get package)
- tb927_3.gff (Gene Transfer annotation file of T. brucei downloaded from TriTrypDB)
- tb927_5.fa (Genomic sequences of *T. brucei* downloaded from TriTrypDB)
- tb927_5.fa.fai (index file Genomic sequences of *T. brucei* downloaded from TriTrypDB)
- tb927_5.gtf (Gene Transfer annotation file of *T. brucei* downloaded from TriTrypDB and supplemented with the new discovered expressed sequences)
- util.py (python script with helper functions for the gene coverage analysis)

wcar.png (Wellcome Centre for Anti-Infectives Research logo)

The QC output is avaiable at github https://github.com/mtinti/polysome_qc, visualizable at https://polysome-qc.onrender.com and archived in zenodo.

Zenodo: mtinti/mtinti-polysome_qc. https://doi.org/10.5281/zenodo.4235212[75]

This project contains the following data:

- report.html (the home page of the visualization report)
- report_data (the configuration folder congaing the report data)

Zenodo: mtinti/polysome_cds[76]:

This project contains the following data:

- (B,P)_(pol, sub, tot)_(1,2,3)
  - counts_CDS.txt (The read counts for the gene coding sequences)

- Figures_Paper_def.ipynb (The jupyter notebook producing the new figure 17 of the paper)

Licence: MIT.

Extended data

Zenodo: v0.3 mtinti/polysome_extended: v0.3 update table 6. https://doi.org/10.5281/zenodo.5884563[77]

This project contains the following extended data:

- **Table 3. Cluster analysis.** Data used for the cluster analysis. The first column reports the gene identification number and 18 columns with the normalized values for the biological replicates in the format of [B/P]_[tot/pol/sub]_[1/2/3] were B: bloodstream form, P: procyclic form, tot: total RNA sample, pol: polysomal sample, sub: subpolysomal sample, 1,2,3: biological replicate identifiers. The table also reports the predicted cluster identification number (label), a binary column reporting whether the gene is identified or not in the (is_ks), the gene description (desc), a binary column reporting whether the gene is annotated as an H/ACA-like snoRNA, a binary column reporting whether the gene is annotated as a snoRNA and a binary column reporting whether the gene is annotated as non-coding (Noncoding) RNA.

- **Table 4. Polysome/sub-polysome transcript differential abundance in BSF cells.** Comparison between the polysome and sub-polysome samples in the bloodstream form life stage: logFC, the log fold-change for each gene in the two groups being compared. logCPM, the log-average abundance for each gene in the two groups being compared. LR, likelihood ratio statistic. PValue, exact p-value for differential expression test. FDR, the p-value adjusted for multiple testing with the Benjamini–Hochberg method (false discovery rate).

- **Table 5. Polysome / Sub-polysome Transcript Differential Abundance in PCF.** Comparison between the polysome and subpolysome samples in the procyclic form life stage the: logFC, the log-abundance ratio, i.e. fold change, for each gene in the two groups being compared; logCPM, the log-average concentration/abundance for each gene in the two groups being compared; LR, likelihood ratio statistics; PValue, exact p-value for differential expression test; FDR, the p-value adjusted for multiple testing with the Benjamini–Hochberg method.

- **Table 6. lncRNA and Surrounding Genes.** Comparison between the polysome and sub-polysome samples for the lncRNAs (gene_ks) and the genes at their 5' (gene_sensitive_at_5prime) or 3' (gene_sensitive_at_3prime) reporting the: logFC, the log fold-change for each lncRNA in the two groups being compared. FDR, the p-value adjusted for multiple testing with the Benjamini–Hochberg method for the lncRNAs. logFC_5p, the log fold-change for the genes at the 5' of the lncRNAs. FDR_5p, the p-value adjusted for multiple testing with the Benjamini–Hochberg method for the genes at the 5' of the lncRNAs. logFC_3p, the log fold-change for the genes at the 3' of the lncRNAs. FDR_3p, the p-value adjusted for multiple testing with the Benjamini–Hochberg method for the genes at the 3' of the lncRNAs. Desc_5p, the gene description for the genes at the 5' of the lncRNAs. Desc_3p, the gene description for the genes at the 3' of the lncRNAs.

- **Table 7. New Protein Coding Genes.** The ID of the new predicted protein coding genes (Gene), the number of peptides identified in mass spectrometry (Peptides found by MS), the genomic coordinates (Coordinates), the gene length in base pairs (Gene length), the open reding frame orientation (Orient), the coding sequence coordinate (CDS coordinates), the open reading frame length in base pairs (ORF length), the predicted protein length in amino acid residues (Predicted protein length), the predicted protein molecular weight in Kilodalton (Predicted protein estimated weight), the identification number of other genes with high homology with the predicted gene (Similar genes), the number of transmembrane domain predicted with the Phobius algorithm (Phobius predictions), the signal peptide prediction results computed with the SignalP 3 algorithm (SignalP 3.0 predictions) or SignalP 5 algorithm (SignalP 5.0 predictions), are reported for the new predicted protein coding gene manually curated.

Acknowledgements

We are grateful to Bernardo Foth for a preliminary analysis of the data. We are grateful to Christine Clayton for helpful comments provided during the preparation of the paper.

References

1. Cox FEG: **History of sleeping sickness (African trypanosomiasis).** *Infect Dis Clin North Am.* 2004; **18**(2): 231–45.
PubMed Abstract | Publisher Full Text

2. Vasquez JJ, Hon CC, Vanselow JT, *et al.*: **Comparative ribosome profiling reveals extensive translational complexity in different *Trypanosoma brucei* life cycle stages.** *Nucleic Acids Res.* 2014; **42**(6): 3623–37.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Johnson PJ, Kooter JM, Borst P: **Inactivation of transcription by UV**

irradiation of T. brucei provides evidence for a multicistronic transcription unit including a VSG gene. *Cell.* 1987; **51**(2): 273–81.
PubMed Abstract | Publisher Full Text

4. Huang J, van der Ploeg LH: **Maturation of polycistronic pre-mRNA in *Trypanosoma brucei*: analysis of trans splicing and poly(A) addition at nascent RNA transcripts from the hsp70 locus.** *Mol Cell Biol.* 1991; **11**(6): 3180–90.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Ullu E, Matthews KR, Tschudi C: **Temporal order of RNA-processing reactions in trypanosomes: rapid trans splicing precedes polyadenylation of newly synthesized tubulin transcripts.** *Mol Cell Biol.* 1993; **13**(1): 720–5.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Boothroyd JC, Cross GA: **Transcripts coding for variant surface glycoproteins of *Trypanosoma brucei* have a short, identical exon at their 5′ end.** *Gene.* 1982; **20**(2): 281–9.
PubMed Abstract | Publisher Full Text

7. Parsons M, Nelson RG, Watkins KP, *et al.*: **Trypanosome mRNAs share a common 5′ spliced leader sequence.** *Cell.* 1984; **38**(1): 309–16.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Van der Ploeg LH, Liu AY, Michels PA, *et al.*: **RNA splicing is required to make the messenger RNA for a variant surface antigen in trypanosomes.** *Nucleic Acids Res.* 1982; **10**(12): 3591–604.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Gilinger G, Bellofatto V: **Trypanosome spliced leader RNA genes contain the first identified RNA polymerase II gene promoter in these organisms.** *Nucleic Acids Res.* 2001; **29**(7): 1556–64.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Sather S, Agabian N: **A 5′ spliced leader is added in trans to both alpha- and beta-tubulin transcripts in Trypanosoma brucei.** *Proc Natl Acad Sci U S A.* 1985; **82**(17): 5695–9.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Jensen BC, Ramasamy G, Vasconcelos EJ, *et al.*: **Extensive stage-regulation of translation revealed by ribosome profiling of *Trypanosoma brucei*.** *BMC Genomics.* 2014; **15**(1): 911.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Antwi EB, Haanstra JR, Ramasamy G, *et al.*: **Integrative analysis of the *Trypanosoma brucei* gene expression cascade predicts differential regulation of mRNA processing and unusual control of ribosomal protein expression.** *BMC Genomics.* 2016; **17**: 306.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Trindade S, Rijo-Ferreira F, Carvalho T, *et al.*: ***Trypanosoma brucei* Parasites Occupy and Functionally Adapt to the Adipose Tissue in Mice.** *Cell Host Microbe.* 2016; **19**(6): 837–48.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Qiu Y, Milanes JE, Jones JA, *et al.*: **Glucose Signaling Is Important for Nutrient Adaptation during Differentiation of Pleomorphic African Trypanosomes.** *mSphere.* 2018; **3**(5): e00366–18.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Archer SK, Inchaustegui D, Queiroz R, *et al.*: **The cell cycle regulated transcriptome of *Trypanosoma brucei*.** *PLoS One.* 2011; **6**(3): e18425.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Capewell P, Monk S, Ivens A, *et al.*: **Regulation of *Trypanosoma brucei* Total and Polysomal mRNA during Development within Its Mammalian Host.** *PLoS One.* 2013; **8**(6): e67069.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Mašek T, Valášek L, Pospíšek M: **Polysome analysis and RNA purification from sucrose gradients.** *Methods Mol Biol.* 2011; **703**: 293–309.
PubMed Abstract | Publisher Full Text

18. Spirin AS: **Ribosome as a molecular machine.** *FEBS Lett.* 2002; **514**(1): 2–10.
PubMed Abstract | Publisher Full Text

19. Pradet-Balade B, Boulmé F, Beug H, *et al.*: **Translation control: bridging the gap between genomics and proteomics?** *Trends Biochem Sci.* 2001; **26**(4): 225–9.
PubMed Abstract | Publisher Full Text

20. Wirtz E, Leal S, Ochatt C, *et al.*: **A tightly regulated inducible expression system for conditional gene knock-outs and dominant-negative genetics in *Trypanosoma brucei*.** *Mol Biochem Parasitol.* 1999; **99**(1): 89–101.
PubMed Abstract | Publisher Full Text

21. Hirumi H, Hirumi K: **Axenic culture of African trypanosome bloodstream forms.** *Parasitol Today.* 1994; **10**(2): 80–4.
PubMed Abstract | Publisher Full Text

22. Brun R, Schönenberger: **Cultivation and *in vitro* cloning or procyclic culture forms of *Trypanosoma brucei* in a semi-defined medium. Short communication.** *Acta Trop.* 1979; **36**(3): 289–92.
PubMed Abstract

23. Aslett M, Aurrecoechea C, Berriman M, *et al.*: **TriTrypDB: a functional genomic resource for the Trypanosomatidae.** *Nucleic Acids Res.* 2010; **38**(Database issue): D457–62.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods.* 2012; **9**(4): 357–9.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–9.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–2.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.** *Bioinformatics.* 2014; **30**(7): 923–30.
PubMed Abstract | Publisher Full Text

28. Siegel TN, Hekstra DR, Wang X, *et al.*: **Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites.** *Nucleic Acids Res.* 2010; **38**(15): 4946–57.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Radío S, Fort RS, Garat B, *et al.*: **UTRme: A Scoring-Based Tool to Annotate Untranslated Regions in Trypanosomatid Genomes.** *Front Genet.* 2018; **9**: 671.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Pertea G, Pertea M: **GFF Utilities: GffRead and GffCompare [version 2; peer review: 3 approved].** *F1000Res.* 2020; **9**: ISCB Comm J-304.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Guegan F, Bento F, Neves D, *et al.*: **A long non-coding RNA controls parasite differentiation in African trypanosomes.** *bioRxiv.* 2020; 2020.05.03.074625.
Publisher Full Text

32. Grabherr MG, Haas BJ, Yassour M, *et al.*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nat Biotechnol.* 2011; **29**(7): 644–52.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Shao M, Kingsford C: **Accurate assembly of transcripts through phase-preserving graph decomposition.** *Nat Biotechnol.* 2017; **35**(12): 1167–1169.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Neph S, Kuehn MS, Reynolds AP, *et al.*: **BEDOPS: high-performance genomic feature operations.** *Bioinformatics.* 2012; **28**(14): 1919–20.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Pertea M, Pertea GM, Antonescu CM, *et al.*: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.** *Nat Biotechnol.* 2015; **33**(3): 290–5.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Wu TD, Watanabe CK: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences.** *Bioinformatics.* 2005; **21**(9): 1859–75.
PubMed Abstract | Publisher Full Text

37. Clark K, Karsch-Mizrachi I, Lipman DJ, *et al.*: **GenBank.** *Nucleic Acids Res.* 2016; **44**(D1): D67–72.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Cock PJ, Antao T, Chang JT, *et al.*: **Biopython: freely available Python tools for computational molecular biology and bioinformatics.** *Bioinformatics.* 2009; **25**(11): 1422–3.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Okonechnikov K, Conesa A, García-Alcalde F: **Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data.** *Bioinformatics.* 2016; **32**(2): 292–4.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–8.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Virtanen P, Gommers R, Oliphant TE, *et al.*: **SciPy 1.0: fundamental algorithms for scientific computing in Python.** *Nat Methods.* 2020; **17**(3): 261–272.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Hansen KD, Irizarry RA, Wu Z: **Removing technical variability in RNA-seq data using conditional quantile normalization.** *Biostatistics.* 2012; **13**(2): 204–16.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Robinson MD, Smyth GK: **Small-sample estimation of negative binomial dispersion, with applications to SAGE data.** *Biostatistics.* 2008; **9**(2): 321–32.
PubMed Abstract | Publisher Full Text

44. McKinney W: **Data structures for statistical computing in python.** *Proceedings of the 9th Python in Science Conference.* 2010; **445**: 56–61.
Publisher Full Text

45. Bengfort B, Bilbro R: **Yellowbrick: Visualizing the Scikit-Learn Model Selection Process.** *J Open Source Softw.* 2019; **4**(35): 1075.
Publisher Full Text

46. Pedregosa F, Varoquaux G, Gramfort A, *et al.*: **Scikit-learn: Machine Learning in Python.** *J Mach Learn Res.* 2011; **12**: 2825–2830.
Reference Source

47. Klopfenstein DV, Zhang L, Pedersen BS, *et al.*: **GOATOOLS: A Python library for Gene Ontology analyses.** *Sci Rep.* 2018; **8**(1): 10872.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Tinti M, Güther MLS, Crozier TWM, *et al.*: **Proteome turnover in the bloodstream and procyclic forms of *Trypanosoma brucei* measured by quantitative proteomics [version 1; peer review: 3 approved].** *Wellcome Open Res.* 2019; **4**: 152.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Käll L, Krogh A, Sonnhammer EL: **A combined transmembrane topology and**

signal peptide prediction method. *J Mol Biol.* 2004; **338**(5): 1027–36.
PubMed Abstract | Publisher Full Text

50. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, *et al.*: **SignalP 5.0 improves signal peptide predictions using deep neural networks.** *Nat Biotechnol.* 2019; **37**(4): 420–423.
PubMed Abstract | Publisher Full Text

51. Camacho C, Coulouris G, Avagyan V, *et al.*: **BLAST+: architecture and applications.** *BMC Bioinformatics.* 2009; **10**: 421.
PubMed Abstract | Publisher Full Text | Free Full Text

52. Madeira F, Park YM, Lee J, *et al.*: **The EMBL-EBI search and sequence analysis tools APIs in 2019.** *Nucleic Acids Res.* 2019; **47**(W1): W636–W641.
PubMed Abstract | Publisher Full Text | Free Full Text

53. Egorov AA, Sakharova EA, Anisimova AS, *et al.*: **svist4get: a simple visualization tool for genomic tracks from sequencing experiments.** *BMC Bioinformatics.* 2019; **20**(1): 113.
PubMed Abstract | Publisher Full Text | Free Full Text

54. Amid C, Alako BTF, Kadhirvelu VB, *et al.*: **The European Nucleotide Archive in 2019.** *Nucleic Acids Res.* 2020; **48**(D1): D70–D76.
PubMed Abstract | Publisher Full Text | Free Full Text

55. Chen S, Zhou Y, Chen Y, *et al.*: **fastp: an ultra-fast all-in-one FASTQ preprocessor.** *Bioinformatics.* 2018; **34**(17): i884–i890.
PubMed Abstract | Publisher Full Text | Free Full Text

56. Stöcklein W, Piepersberg W: **Binding of cycloheximide to ribosomes from wild-type and mutant strains of *Saccharomyces cerevisiae*.** *Antimicrob Agents Chemother.* 1980; **18**(6): 863–7.
PubMed Abstract | Publisher Full Text | Free Full Text

57. Gopal S, Cross GA, Gaasterland T: **An organism-specific method to rank predicted coding regions in *Trypanosoma brucei*.** *Nucleic Acids Res.* 2003; **31**(20): 5877–85.
PubMed Abstract | Publisher Full Text | Free Full Text

58. Risso D, Schwartz K, Sherlock G, *et al.*: **GC-content normalization for RNA-Seq data.** *BMC Bioinformatics.* 2011; **12**: 480.
PubMed Abstract | Publisher Full Text | Free Full Text

59. Mandelboum S, Manber Z, Elroy-Stein O, *et al.*: **Recurrent functional misinterpretation of RNA-seq data caused by sample-specific gene length bias.** *PLoS Biol.* 2019; **17**(11): e3000481.
PubMed Abstract | Publisher Full Text | Free Full Text

60. mtinti: **mtinti/polysome_extended: v0.1 (Version v0.2)**. *Zenodo*. 2021.
http://www.doi.org/10.5281/zenodo.4526335

61. Mugo E, Clayton C: **Expression of the RNA-binding protein RBP10 promotes the bloodstream-form differentiation state in *Trypanosoma brucei*.** *PLoS Pathog.* 2017; **13**(8): e1006560.
PubMed Abstract | Publisher Full Text | Free Full Text

62. Rico E, Ivens A, Glover L, *et al.*: **Genome-wide RNAi selection identifies a regulator of transmission stage-enriched gene families and cell-type differentiation in *Trypanosoma brucei*.** *PLoS Pathog.* 2017; **13**(3): e1006279.
PubMed Abstract | Publisher Full Text | Free Full Text

63. Erben ED, Fadda A, Lueong S, *et al.*: **A genome-wide tethering screen reveals novel potential post-transcriptional regulators in *Trypanosoma brucei*.** *PLoS Pathog.* 2014; **10**(6): e1004178.
PubMed Abstract | Publisher Full Text | Free Full Text

64. Kolev NG, Franklin JB, Carmi S, *et al.*: **The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution.** *PLoS Pathog.* 2010; **6**(9): e1001090.
PubMed Abstract | Publisher Full Text | Free Full Text

65. Carlevaro-Fita J, Rahim A, Guigó R, *et al.*: **Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells.** *RNA.* 2016; **22**(6): 867–82.
PubMed Abstract | Publisher Full Text | Free Full Text

66. Statello L, Guo CJ, Chen LL, *et al.*: **Gene regulation by long non-coding RNAs and its biological functions.** *Nat Rev Mol Cell Biol.* 2021; **22**(2): 96–118.
PubMed Abstract | Publisher Full Text | Free Full Text

67. Liang XH, Uliel S, Hury A, *et al.*: **A genome-wide analysis of C/D and H/ACA-like small nucleolar RNAs in *Trypanosoma brucei* reveals a trypanosome-specific pattern of rRNA modification.** *RNA.* 2005; **11**(5): 619–45.
PubMed Abstract | Publisher Full Text | Free Full Text

68. Chikne V, Shanmugha Rajan K, Shalev-Benami M, *et al.*: **Small nucleolar RNAs controlling rRNA processing in *Trypanosoma brucei*.** *Nucleic Acids Res.* 2019; **47**(5): 2609–2629.
PubMed Abstract | Publisher Full Text | Free Full Text

69. Reinisch KM, Wolin SL: **Emerging themes in non-coding RNA quality control.** *Curr Opin Struct Biol.* 2007; **17**(2): 209–14.
PubMed Abstract | Publisher Full Text

70. Slomovic S, Laufer D, Geiger D, *et al.*: **Polyadenylation of ribosomal RNA in human cells.** *Nucleic Acids Res.* 2006; **34**(10): 2966–75.
PubMed Abstract | Publisher Full Text | Free Full Text

71. Martin G, Keller W: **RNA-specific ribonucleotidyl transferases.** *RNA.* 2007; **13**(11): 1834–49.
PubMed Abstract | Publisher Full Text | Free Full Text

72. Leinonen R, Sugawara H, Shumway M, *et al.*: **The sequence read archive.** *Nucleic Acids Res.* 2011; **39**(Database issue): D19–21.
PubMed Abstract | Publisher Full Text | Free Full Text

73. mtinti: **mtinti/polysome: pre-submission (Version v1.1)**. *Zenodo*. 2021.
http://www.doi.org/10.5281/zenodo.4447412

74. mtinti: **mtinti/polysome_coverage: pre-submissiom (Version v1.0)**. *Zenodo*. 2021.
http://www.doi.org/10.5281/zenodo.4447015

75. mtinti: **mtinti/polysome_qc: activate zenodo (Version 0.1)**. *Zenodo*. 2020.
http://www.doi.org/10.5281/zenodo.4235213

76. mtinti: **mtinti/polysome_cds: 0.1.** *Zenodo.* 2022.
http://www.doi.org/10.5281/zenodo.5886850

77. mtinti: **mtinti/polysome_extended: v0.3 update table 6.** *Zenodo.* 2022.
http://www.doi.org/10.5281/zenodo.5884563

Open Peer Review

Current Peer Review Status: ? ✓ ✓ ✓

Version 3

Reviewer Report 15 February 2022

https://doi.org/10.21956/wellcomeopenres.19517.r48406

✓ **Osvaldo P de Melo Neto**
Department of Microbiology, Instituto Aggeu Magalhães, Fiocruz, Recife, PE, 50740-465, Brazil
Antonio Rezende
Department of Microbiology, Instituto Aggeu Magalhães, Fiocruz, Recife, Brazil

I have seen the revised version of the manuscript and the replies to the issues and comments raised by me and my colleague Antonio Rezende. I can confirm that I am satisfied with the answers and modifications carried out by the authors in the revised version and that I can fully "Approve" the manuscript without reservations.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Molecular Biology, Protozoology, Bioinformatics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 08 February 2022

https://doi.org/10.21956/wellcomeopenres.19517.r48404

✓ **Esteban Erben**
IIBIO-UNSAM, Buenos Aires, Argentina

The authors have satisfactorily addressed most of my concerns.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Molecular Parasitology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 05 July 2021

https://doi.org/10.21956/wellcomeopenres.18079.r44178

?

Osvaldo P de Melo Neto
Department of Microbiology, Instituto Aggeu Magalhães, Fiocruz, Recife, PE, 50740-465, Brazil
Antonio Rezende
Department of Microbiology, Instituto Aggeu Magalhães, Fiocruz, Recife, Brazil

The manuscript by Tinti *et al.* sought to generate profiles of mRNAs associated with polysomal and non-polysomal fractions from the parasitic protozoa “Trypanosoma brucei”; polysomal mRNAs indicating those being actively translated and used for protein synthesis. The study compares the profiles seen for the two major life forms of “T. brucei” (bloodstream and procyclic) using as biological material fractions of cycloheximide-treated cells. Pooled samples consisting of polysomal and non-polysomal fractions, separated by sucrose gradient, were submitted to RNA extraction, cDNA synthesis and next generation sequencing in order to qualitatively and quantitatively define mRNAs (or polyadenylated RNAs) bound or not bound to the polysomes. In all, the authors investigated three different samples for each parasite life form: total polyA+ mRNAs, sub-polysomal mRNAs and polysomal mRNAs. A large range of bioinformatic tools were used to analyze the results and these were also compared with previous published data.

The data generated by the reported research, with the comparison between polysomal and sub-polysomal mRNAs from the two different life forms, expands on previous works assessing polysome bound mRNAs in *T. brucei*, increases the number of experimentally validated and mapped mRNAs and identifies and investigates non-coding mRNAs in the sub-polysomal fraction, with possible roles in regulation of specific messages. It constitutes an important tool for a large number of investigators working on mRNA translation and gene expression regulation in these and related protozoa. The approach is especially relevant considering that these are eukaryotes with very peculiar mechanisms for regulation of gene expression, with previous evidence highlighting a strong role for this regulation during mRNA translation, therefore being model organisms for mechanisms targeting regulation of translation. Nevertheless, for publication, the manuscript needs improvements regarding several issues, as detailed below.

Abstract:

The abstract should be thoroughly revised with an inclusion of more relevant results. As detailed further below, a focus only on the presence of long non-coding RNA in the sub-polysomal fraction (stated twice) is misleading since these should be mainly found in this fraction to begin with. More relevant results include: the proper identification/confirmation of long non-coding, polyadenylated RNAs (with the 5' SL), with substantial number of reads and their mapping, highlighting possible roles regarding the regulation of specific protein coding genes (it is not clear by the text to which extent novel lncRNAs were identified, but if so this has to be highlighted as well as the association reported between lcnRNAs and neighboring protein coding transcripts); the substantial increase of genome loci found to be transcribed, as reported in the main text; and the identification of novel hypothetical proteins.

Methods:

The approach applied in the manuscript was based on several bioinformatic tools and computational steps. Some of the tools used are quite old, such as GMAP, but more relevant was the use of several computational tools and analyses without a proper justification for their use. For example, why did the authors decide to use two different *de novo* assemblers (Trinity and Scallop) for the transcriptomic data? A consensus between the transcript predictions from both tools and including prior predictions was then expected, but it is not clear if this is what the authors did and should be clarified. At the end of the "de novo" approach, a new genome annotation file was generated, but it is not clear if it was used for the final read counts using the "featureCounts" tool and this also needs clarification. Regarding the description of the cDNA library preparation, important details such as read length and which type of library was employed cannot be found in the methods section. Base on the text it is understood that the libraries used were *paired-end*, but then for quality control, the authors mentioned they used RPKM as normalization step, which is applicable to *single-end* libraries. These details need to be better clarified.

An important point of concern is the use of two different strains, one for each of the two *T. brucei* life forms. No considerations or comments are made on how this can impact on the differences found between the same fractions from the two different life forms investigated. To what extent could the differences seen between the two life forms, and related to the results from figure 5, could be associated with the use of different strains?

Results:

The very large number of figures can be a distraction and keep the reader from focusing on what is important. The authors should consider reducing those. Some suggestions: Figures 1 is not necessary; for Figures 6 and 7, only one representative figure could be kept, or both could be removed altogether; Figure 9 also does not need to be shown.

For validation purposes, wouldn´t it be relevant mentioning or showing the profile of known, stage specific genes, which would preferentially be present in the polysomal fractions of either bloodstream or procyclic forms? For instance, known surface antigens?

The authors used as examples on mapping quality control, two *T. brucei* genes known to have introns. However, it is not clear why one should see substantial read coverage for an intron in sub-polysomal and total RNA samples. It was understood that the work aimed to compare sub-

polysomal or polysomal fractions of presumably cytoplasmic and mature mRNAs, selected through their poly-A tails. So, very low or no intron mapping would be expected overall and not only for the polysomal fractions. Wouldn't the presence of introns indicate a substantial amount of precursor or maybe nuclear mRNAs within the sub-polysomal fractions that need to be considered somehow?

A relevant observation from the results shown in Figure 4 is the overlap in the UTRs from the Tb927.8.1500 and Tb927.8.1510. How unique is this? Likewise, in Figure 3, for the Tb927.3.3160 gene, its intron might be associated with poly-A tracts. Again, what does this mean and is it seen elsewhere? These issues should be considered in the text.

The authors mention an enrichment of lncRNAs in cluster 2, composed mainly by sub-polysomal mRNAs from bloodstream and procyclic forms. Biologically speaking, isn't that to be expected, as no ribosomes should be attached to non-translatable lncRNAs? The authors propose a mechanism where a lncRNA would be part of a transcript, and its presence or absence could define different transcript isoforms. But which isoform would be translated, with or without the lncRNA segment? In this case would the lncRNA segment be considered a long non-coding RNA? Why would the lncRNA have a Spliced Leader and poly-A, as indicated by the results? These issues were discussed only superficially in the Discussion.

The authors also proposes that the location of the segments encoding several lncRNA might indicate a role in regulation of neighboring genes. What is the basis for this? Has it been shown elsewhere or in Tryps? It seems quite speculative and not much related to this is seen in the discussion.

Finally, it would be nice to have some data regarding the abundance of different functional classes of mRNA coding proteins when forms and fractions were compared, however, there is nothing on that in the discussion.

Discussion:

Regarding the discussion, the work has several interesting results, but few of them are discussed sufficiently. For instance, the results mention a 34% increment in the number of transcribed loci, but there is no consideration regarding this very relevant results in the discussion, and no details can be found regarding function and curation of these loci.

Minor Points:

1. What is the difference between PCA and MDS? Do the axes have weights? This can change the interpretation of the result. Why didn't authors use PCA?

2. In the methods, in the sentence "Before computing the fraction of transcripts in polysomes, the polysome read counts were divided by 0.7 and the sub-polysome read counts were divided by 0.3 …", shouldn't the counts be multiplied by 0.7 or 0.3, instead of divided by?

3. The following sentence is not understandable and needs to be clarified: "To identify the Grumpy Like genes we created a third model to study the differential transcript abundance between the sub-polysomal samples (mixed model of BSF and PCF) against the subpolysomal samples (mixed model of BSF and PCF)" present at "Sub-polysome / polysome

differential abundance analysis and Grumpy Like GenesSub-polysome / polysome differential abundance analysis and Grumpy Like Genes" topic.

4. The first sentences of the Results section explaining in detail the use and properties of cycloheximide are not necessary and can be shortened since the cycloheximide use for polysomal gradients is common.

5. The statement “The poly(A) genomic tracts are often present in intergenic regions and can help to determine the 3’ gene boundaries” needs a reference.

6. In various Figure legends the citation of Vasquex (it is Vasquez) *et al.* 2014 should include the reference number from the reference list.

7. Figure 8 - put the legend of the figure in the same orders of the sample names in the graph.

8. For Figure 10, review the order that the results are shown. It makes no sense having the two different sub-polysomal samples followed by the polysomal and total samples from procyclics and then polysomal and total samples from bloodstream. Either you alternate between each of the procyclic and bloodstream samples or show all procyclic and then all bloodstream samples.

9. The following sentence needs to be modified “The grumpy transcript made us wonder which other sub-polysome enriched transcripts might have a lncRNA at the 5’ end….”. Based on the “Grumpy” example, it refers to polysome enriched transcripts which have a sub-polysome enriched lncRNA at their 5’ end.

10. Figure 13 - The legend is confusing.

11. In the sentence "Synteny analysis of the TRY.375 locus performed at TryTripDB revealed another gene (TevSTIB805.7.3380) in the T. evansi genome with 100% homology with the predicted TRY.375 gene product.", the homology concept is misused as it is a binary concept bringing the idea of evolutionary relationship, so you cannot have degree of homology between two loci of different species. Either they are homologs or re not. The degree of similarity or identity should be informed.

12. The novel MSTRG.94 gene and neighboring sequences seem to be transcribed only in Bloodstream cells and this should be highlighted.

13. In the Discussion, the statement “This is demonstrated by the virtual absence of reads covering the intron regions of the two experimentally validated intron containing genes” is not valid for the sub-polysomal transcripts and should be revised.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Molecular Biology of Prozooans; Gene expression regulation

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 21 Jan 2022

Michele Tinti, School of Life Sciences, University of Dundee, Dundee, Dundee, UK

1. The very large number of figures can be a distraction and keep the reader from focusing on what is important.

We agree that we have many figures and most of them would be more appropriate as supplementary figures in other journals. However, as the WOR publication allows “as many figure as needed”, we decided to include a visual support for many of our statements.

1. For example, why did the authors decide to use two different de novo assemblers (Trinity and Scallop) for the transcriptomic data?

We found Trinity and Scallop identify different sets of transcribed regions. However, both assemblers were developed with eukaryotic genes with introns and struggled to correctly identify new genes in *T. brucei*. Particularly Trinity was prone to assemble transcripts encompassing several genes. For this reason, we run Scallop first to annotate new transcripts in regions without any previous annotation. Then, we repeated the same analysis with Trinity, again considering only regions without previous annotation. We have added an explanation to the paper materials and methods.

1. The authors mentioned they used RPKM as normalization step.

This was a mistake; it should have read as FPKM, and this has been corrected.

1. An important point of concern is the use of two different strains, one for each of the two T. brucei life forms. No considerations or comments are made on how this can impact on the differences found between the same fractions from the two different life forms investigated. To what extent could the differences seen between the two life forms, and related to the results from figure 5, could be associated with the use of different strains?

The data we produced form procyclic and bloodstream forms both come from the Lister 427 strain. This is now described more explicitly in Material and Methods.

1. For validation purposes, wouldn´t it be relevant mentioning or showing the profile of known, stage specific genes, which would preferentially be present in the polysomal fractions of either bloodstream

We did check a few key proteins that are bloodstream or procyclic life stage specific. Also, In Fig 5, we are already showing that our data is in good agreement with previous results (this has been also revisited as suggested by the reviewer in the minor point n 2).

1. However, it is not clear why one should see substantial read coverage for an intron in sub-polysomal and total RNA samples. It was understood that the work aimed to compare sub-polysomal or polysomal fractions of presumably cytoplasmic and mature mRNAs, selected through their poly-A tails.

We believe that the sub-polysomal and total RNA fractions contain transcripts that for some reasons failed the splicing event and are targeted for degradation. We would expect these failed splicing events to be virtually absent from the actively transcribing polysomal fraction.

1. A relevant observation from the results shown in Figure 4 is the overlap in the UTRs from the Tb927.8.1500 and Tb927.8.1510. How unique is this? Likewise, in Figure 3, for the Tb927.3.3160 gene, its intron might be associated with poly-A tracts. Again, what does this mean and is it seen elsewhere? These issues should be considered in the text.

We address the difficulties to determine the 3’ UTRs end genomic coordinates in the new version of the article. We also reanalysed our data to consider only the lncRNAs and the CDS of protein coding genes for our statistical analysis. (see new chapter: Expression Analysis of lncRNAs and surrounding genes, Figure 17 and associated text).

1. The authors mention an enrichment of lncRNAs in cluster 2, composed mainly by sub-polysomal mRNAs from bloodstream and procyclic forms. Biologically speaking, isn’t that to be expected, as no ribosomes should be attached to non-translatable

lncRNAs?

We have revised the discussion text highlighting a possible role for lncRNAs more abundant in the polysomal fraction.

1. The authors propose a mechanism where a lncRNA would be part of a transcript, and its presence or absence could define different transcript isoforms. But which isoform would be translated, with or without the lncRNA segment? In this case would the lncRNA segment be considered a long non-coding RNA? Why would the lncRNA have a Spliced Leader and poly-A, as indicated by the results? These issues were discussed only superficially in the Discussion.

We agree with the reviewer that no experimental evidence supports this speculative hypothesis and we have modified this part of the discussion.

1. The authors also proposes that the location of the segments encoding several lncRNA might indicate a role in regulation of neighboring genes. What is the basis for this? Has it been shown elsewhere or in Tryps? It seems quite speculative and not much related to this is seen in the discussion.

We will add a reference to “Statello et al. 2021, Nature Reviews Molecular Cell Biology. Gene regulation by long non-coding RNAs and its biological functions” providing pieces of evidence that “Several lncRNAs control the expression of nearby genes by affecting their transcription”. We also provided more robust statistical analysis to correlate the expression of the lncRNAs and the genes at their 5’ or 3’

1. Regarding the discussion, the work has several interesting results, but few of them are discussed sufficiently. For instance, the results mention a 34% increment in the number of transcribed loci, but there is no consideration regarding this very relevant results in the discussion, and no details can be found regarding function and curation of these loci.

We have clarified that the majority of those loci represent miss-annotated UTRs. This is the reason why we used an unbiased proteomic approach to shortlist putative new protein-coding genes. It would be beyond our capacity to manually curate all the identified new loci and to better annotate the 3’ UTRs of many genes.

Minor Points

1. What is the difference between PCA and MDS? Do the axes have weights? This can change the interpretation of the result. Why didn’t authors use PCA?

We prefer the MDS analysis as the axes have a direct correlation with the differential expression of the considered genes, see https://www.biostars.org/p/287415/.

1. In the methods, in the sentence “Before computing the fraction of transcripts in polysomes, the polysome read counts were divided by 0.7 and the sub-polysome read counts were divided by 0.3 ...”, shouldn’t the counts be multiplied by 0.7 or 0.3, instead of divided by?

The reviewer is correct, and this improves the correlation between our dataset and the dataset of Antwi et al. showed in Figure 5

1. The following sentence is not understandable and needs to be clarified: "To identify the Grumpy Like genes we created a third model to study the differential transcript abundance between the sub-polysomal samples (mixed model of BSF and PCF) against the subpolysomal samples (mixed model of BSF and PCF)" present at "Sub-polysome / polysome differential abundance analysis and Grumpy Like GenesSub-polysome / polysome differential abundance analysis and Grumpy Like Genes" topic.

We apologise for the poor construction and have clarified in the revised version (see new chapter: Expression Analysis of lncRNAs and surrounding genes, Figure 17 and associated text).[MF(5] [MT(6]

1. The first sentences of the Results section explaining in detail the use and properties of cycloheximide are not necessary and can be shortened since the cycloheximide use for polysomal gradients is common.

We prefer to keep this sentence for people not familiar with this technique.

1. The statement “The poly(A) genomic tracts are often present in intergenic regions and can help to determine the 3’ gene boundaries” needs a reference.

We have added “ Radío et al. 2018 UTRme: A Scoring-Based Tool to Annotate Untranslated Regions in Trypanosomatid Genomes” to this sentence.

1. In various Figure legends the citation of Vasquex (it is Vasquez) et al. 2014 should include the reference number from the reference list.

We have fixed this.

1. Figure 8 - put the legend of the figure in the same orders of the sample names in the graph.

We have reordered the legend.

1. For Figure 10, review the order that the results are shown. It makes no sense having the two different sub-polysomal samples followed by the polysomal and total samples from procyclics and then polysomal and total samples from bloodstream.

Either you alternate between each of the procyclic and bloodstream samples or show all procyclic and then all bloodstream samples.

This is the order determined by the column clustering, that we would rather keep as it is. We acknowledge that we did not describe the clustering of the columns that is now added to the materials and methods.

1. The following sentence needs to be modified “The grumpy transcript made us wonder which other sub-polysome enriched transcripts might have a lncRNA at the 5’ end….”. Based on the “Grumpy” example, it refers to polysome enriched transcripts which have a sub-polysome enriched lncRNA at their 5’ end.

We have rephrased this sentence.

1. Figure 13 - The legend is confusing.

We don’t find it confusing, we cordially ask the reviewer to specify better where the confusion arises.

1. In the sentence "Synteny analysis of the TRY.375 locus performed at TryTripDB revealed another gene (TevSTIB805.7.3380) in the T. evansi genome with 100% homology with the predicted TRY.375 gene product.", the homology concept is misused as it is a binary concept bringing the idea of evolutionary relationship, so you cannot have degree of homology between two loci of different species. Either they are homologs or re not. The degree of similarity or identity should be informed.

We now report on the degree of similarity.

1. The novel MSTRG.94 gene and neighboring sequences seem to be transcribed only in Bloodstream cells and this should be highlighted.

We have highlighted this.

1. In the Discussion, the statement “This is demonstrated by the virtual absence of reads covering the intron regions of the two experimentally validated intron containing genes” is not valid for the sub-polysomal transcripts and should be revised.

We have rephrased this sentence, referring only to the polysomal samples.

Competing Interests: No competing interests were disclosed.

Reviewer Report 28 June 2021

https://doi.org/10.21956/wellcomeopenres.18079.r44183

Magdalena Radwanska
[1] Biomedical Research Center (BMRC), Ghent University Global Campus, Incheon, South Korea
[2] Ghent University, Ghent, Belgium

This is a high-quality manuscript with sound methodology. The manuscript contributes to the understanding of mechanisms involved in regulation of a gene expression in *T. brucei* using cultured bloodstream forms and procyclic forms. Conducted bioinformatic analysis of RNS-seq on total, sub-polysomal and polysomal mRNA samples led to identification of several long non-coding RNAs (lncRNAs) and snoRNAs implicated in regulation of differentiation in *T. brucei*. In addition presented work includes also identification of 30 new hypothetical protein-coding genes facilitating further genomic annotation of *T. brucei.*
In the host, *T. brucei* as an extracellular parasite surviving in direct contact with the immune system being confronted continuously with the host derived molecules. The latter were shown to have various impacts on parasite survival and differentiation. Hence, the parasite and host interplay and sensing of the environment play an important part in establishing of the infection. In order to make this manuscript more appealing to the wider public, it would be beneficial that authors discuss whether similar regulatory mechanisms involved in gene expression, are expected to operate during real in-host infection. Can observed *in vitro* regulatory mechanisms be extrapolated to the once existing during infection in the context of parasite differentiation, quorum sensing and overall survival?

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Molecular Biology and Immunoparasitology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 11 June 2021

https://doi.org/10.21956/wellcomeopenres.18079.r44184

?

Esteban Erben
IIBIO-UNSAM, Buenos Aires, Argentina

In Tinti and co-workers, the authors look for novel coding sequences and potential lncRNAs by sequencing total RNA and different polysomal fractions from both PCF and BSF-trypanosome forms. The authors found that the grumpy lncRNA (the only lncRNA from *T. brucei* partially characterised so far potentially involved in the regulation of the slender-to-stumpy differentiation) is enriched in subpolysomal fractions in both life forms. They then looked for new lncRNAs displaying similar polysomal distribution finding a bunch of them associated or in close proximity to genes involved in stumpy formation or post-transcriptional regulators suggesting an interesting and yet unexplored functional link between them. If this holds true, it would also mean that tagging some specific proteins from the endogenous locus - a popular approach in this field - (where UTRs are disrupted) may have unforeseen functional consequences. The authors also identified 11 novel protein coding genes. Since both NGS and mass spectrometry rely heavily on databases, I find this manuscript a helpful contribution to the continuous curation of the *T brucei* genome. I also celebrate how detailed the Methods section is. I am supportive of publication, though make some suggestions for possible improvement of the manuscript.

Major criticism:

1. lncRNAs:
   The author found that several lncRNA are in close proximity (typically at the 5'-end) to genes associated with the transition between the BSF and PCF life stages or potential post-transcriptional regulators. Is this statically significant? How does it compare to different functional gene categories or among developmentally vs non-developmentally-regulated genes? The authors show the FC between different fractions but I would like to see how levels of the protein-coding gene and its proximal lncRNA vary in PF vs BF. For instance, what is going on with the lncRNA immediately upstream of the RBP10 gene in PF? Since it seems to be embedded into the 3'UTR of the upstream gene (Tb927.8.2770), is its level independently regulated or follow the Tb927.8.2770 pattern? Most of the lncRNAs are found in subpolysomal fractions. For the nuclear grumpy it is expected. However, RBP7A seems to be cytosolic (also RBP10 for which a subpolysomal lncRNA is apparently linked), so how does the author imagine the crosstalk between the coding genes and their proximal lncRNAs?

2. In the Discussion, the author hypothesized that the excised lncRNA might be targeted for degradation, and this could be the reason why the lncRNAs are enriched in the sub-polysomal fractions. Fig 15 shows that the lncRNA present upstream of the RBP10 coding

region is indeed detected in the ribosome profiling data in both PF and BF stages (although no peptides detected). In human cells, cytoplasmic lncRNAs may indeed be recruited to ribosomes for degradation (PMID: 27090285).[1] Can this also apply for *T. brucei* and what is observed are lncRNA that are for instance either nuclear (like grumpy) or on the way to ribosomes? The authors should discuss all the possible options and do comment about the presence of lncRNA-derived reads on ribosome profiling data.

3. Novel proteins:
   Are the novel proteins developmentally regulated? Could the authors go back to the RITseq from Alsford et al. and check whether those novel genes are required for proliferation?

Minor issues:

1. Is Vasquez *et al.*; not Vasquex *et al.*

2. The reference labelled as #7 may be the #12?

3. In the Discussion, the authors claim that grumpy regulates the transformation from slender to stumpy forms; although the cited reference presents very interesting data, it is still preliminary. I would rephrase it in order to stress it down

References

1. Carlevaro-Fita J, Rahim A, Guigó R, Vardy LA, et al.: Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells.*RNA*. **22** (6): 867-82 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Molecular Parasitology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 21 Jan 2022

Michele Tinti, School of Life Sciences, University of Dundee, Dundee, Dundee, UK

We would like to thank the reviewer for their time giving feedback and providing constructive criticisms to the paper. Here we address the points raised in the reviewer comments. These are also reflected in the revised version of the manuscript.

1. The author found that several lncRNA are in close proximity (typically at the 5'-end) to genes associated with the transition between the BSF and PCF life stages or potential post-transcriptional regulators. Is this statically significant?

We acknowledge that we should determine the statistical significance (if any) of lncRNA proximity to genes known to be involved in post transcriptional regulation. To this end, we performed GO term enrichment analysis of the genes at the 5' of the lncRNA (see new chapter: Expression Analysis of lncRNAs and surrounding genes, Figure 17 and associated text).[MF(1] [MT(2]

1. How does it compare to different functional gene categories or among developmentally vs non-developmentally-regulated genes?

Guegan et al., 2020 reported that 19 *T. brucei* lncRNAs genes were located immediately upstream or downstream of 18 of the 43 SIF pathway genes. As discussed above, we have revised [MF(3] [MT(4] the paper adding a more robust statistical analysis and looking at GO enrichment analysis (see new chapter: Expression Analysis of lncRNAs and surrounding genes, Figure 17 and associated text).[MF(5] [MT(6] .

1. The authors show the FC between different fractions but I would like to see how levels of the protein-coding gene and its proximal lncRNA vary in PF vs BF.

This is an interesting point. However, several lncRNA overlap with the mRNA transcripts. To properly answer this interesting point, we re-mapped our data considering only lncRNAs and the CDS of protein coding genes. We then added more robust statistical analysis of the lncRNAs and the genes at their 5' or 3' (see new chapter: Expression Analysis of lncRNAs and surrounding genes, Figure 17 and associated text).

1. so how does the author imagine the crosstalk between the coding genes and their proximal lncRNAs?

We do not have a working hypothesis at the moment, and we make this clear in the revision.

1. Are the novel proteins developmentally regulated? Could the authors go back to the RITseq from Alsford et al. and check whether those novel genes are required for proliferation?

With respect, we think that these points go beyond the scope of our analysis. We mostly just wanted to flag that a few more putative protein-coding genes exist to the trypanosome research community so that individuals can ask these types of questions .

Minor issues:

1. Is Vasquez et al.; not Vasquex et al.

We correct the misspelling.

1. The reference labelled as #7 may be the #12?

We correct this miss annotation.

1. In the Discussion, the authors claim that grumpy regulates the transformation from slender to stumpy forms; although the cited reference presents very interesting data, it is still preliminary. I would rephrase it in order to stress it down.

We have changed the text to address this issue

Competing Interests: No competing interests were disclosed.

Reviewer Report 15 March 2021

https://doi.org/10.21956/wellcomeopenres.18079.r42787

?

Shulamit Michaeli
The Mina and Everard Goodman Faculty of Life Sciences and Advanced Materials and Nanotechnology Institute, Bar-Ilan University, Ramat-Gan, Israel

The paper by Tinti *et al.* determined the RNA present in polysomal and sub-polysomal fractions by RNA-seq in the two life stages of the trypanosome parasites, PCF and BSF. The study compared their data to published ribosome-profilling and to the polysomal mRNA described by the Clayton group. This is an important study that detects interesting 5' lncRNA. They propose an interesting hypothesis that part of long transcripts are excised out and moving part of mRNA to the sub-polysomal mRNA. They also detected snoRNA in the sub-polysomal fractions and their

interpretation is that these are precursors en-route to be degraded. In addition, they describe new protein coding genes.

Major criticism
The authors suggested based on the paper present in Archives that the 5' lncRNA regulates the downstream gene (ref 31). However, even in the cited paper, there is no direct evidence that the lncRNA regulates the neighboring gene (RBP7A and B genes). There is no evidence that the ncRNA affects the expression of the neighboring genes also in this stud. In the cases presented in Fig. 14, 15, and 16 and especially in Fig. 14, the CDS and the upstream lncRNA are in my eyes individually trans-spliced and polyadenylated transcripts.
So these could be independent entities. The lncRNA may even regulate other transcripts. To convince that there is a continuous transcript between the lncRNA and the downstream gene it is needed to amplify an RNA with primers coming from both genes that is trans-spliced and polyadenylated. Without this experiment, one can not say that the 5' UTR lncRNA is processed from a longer transcript carrying the CDS.
Regarding the snoRNAs, snoRNA are indeed processed from transcripts that are trans-spliced and polyadenylated but are processed by endonucleolytic cleavage. There is no presentation of the snoRNA transcript reads to see the sequence of the intergenic regions (between the snoRNAs) so is not clear to me what the authors refer to i.e. snoRNA precursors or just snoRNA associated with RNP complexes. The association of snoRNAs with complexes from 90 to 60S is expected because snoRNAs are involved in ribosome processing and modification and hence are found in large processing complexes that are smaller in size compared to polysomes and are special processing complexes.

The issues presented above need to be clarified before this paper can be published.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Molecular biology and RNA Biology of trypanosomes

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 21 Jan 2022

Michele Tinti, School of Life Sciences, University of Dundee, Dundee, Dundee, UK

We would like to thank the reviewer for their time giving feedback and providing constructive criticisms to the paper. Here we address the points raised in the reviewer comments. These are also reflected in the revised version of the manuscript.

1. We agree with the referee when she states that “So these (lncRNAs) could be independent entities. The lncRNA may even regulate other transcripts”. In our paper, tried to be cautious about over-stating this as a working hypothesis. However, we thought it worth flagging to the scientific community the proximity of lncRNAs to genes regulating the life stage transitions in T. brucei and thus speculated on their possible roles in regulating gene transcription. In the revised version, we have removed this hypothesis.
2. The reviewer points out that “it is not clear to me what the authors refer to snoRNA precursors or just snoRNA associated with RNP complexes”. We referred to gene annotated as snoRNA in TriTrypDB. It would be hard to discriminate wherever they come from, if precursor or snoRNA associated with RNP complexes.

Competing Interests: No competing interests were disclosed.