Stranded short nascent strand sequencing reveals the topology of eukaryotic DNA replication origins in *Trypanosoma brucei*.

STANOJCIC Slavica [1*#], BARCKMANN Bridlin*[1], MONSIEURS Pieter [2], CROBU Lucien [1], GEORGE Simon [3] and STERKERS Yvon [1#]

[1] University of Montpellier, CNRS, IRD, Academic Hospital (CHU) of Montpellier, MiVEGEC, Montpellier, France.

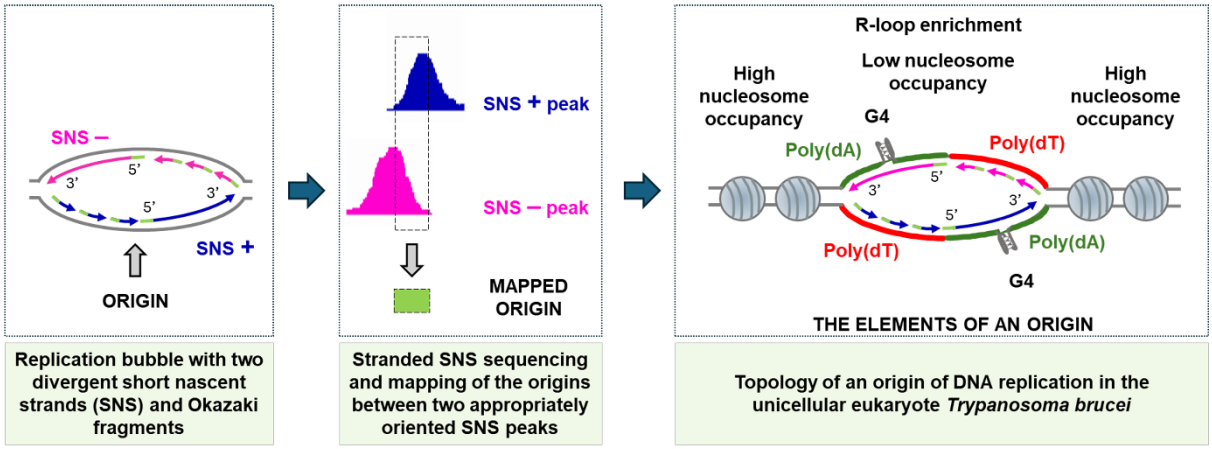
[2] Belgian Nuclear Research Centre (SCK•CEN), Mol, Antwerp, Belgium.

[3] MGX-Montpellier GenomiX, University of Montpellier, CNRS, INSERM, Montpellier, France.

* Authors contributed to the manuscript equally

[#] Corresponding authors: slavica.stanojcic@umontpellier.fr, yvon.sterkers@umontpellier.fr

GRAPHICAL ABSTRACT



ABSTRACT

The structure of eukaryotic origins is a long-standing question in the DNA replication field, and the universal features that define the genomic regions acting as replication starting sites remain unclear. In this study, we employed the stranded SNS-seq methodology, a high-throughput sequencing method that preserves the directionality of short nascent strands, to map a set of origins in *Trypanosoma brucei*. This approach enhanced the specificity and resolution of origin mapping. Our findings indicated that replication predominantly initiates in intergenic regions, situated between poly(dA) and poly(dT) enriched sequences. Experimentally detected G4 structures were observed in the vicinity of some origins. These structures were observed to be embedded within poly(dA) enriched sequences in a strand-specific manner, with the G4s on the plus strand located upstream and the G4s on the minus strand located downstream of the centre. The centre of origin was mainly characterised by low nucleosome occupancy, with flanking regions displaying high nucleosome occupancy. Additionally, our findings revealed that 90% of origins overlapped with previously reported R-loops. To gain further insight, DNA

combing analysis was employed to compare replication at the single-cell level with that observed at the entire cell population level. Finally, a model of eukaryotic origin has been proposed.

Key words
Trypanosoma brucei, DNA replication, origins, replication start sites, Stranded Short Nascent Strands sequencing, stranded SNS-seq, origin clustering, poly(dA), poly(dT) sequences, G4, G-quadruplex, nucleosome depleted regions, R-loops, Okazaki fragments.

Abbreviations

ACS – ARS consensus sequences

ARS – autonomously replicating sequence.

BSF – bloodstream form

G4 – G-quadruplex

gDNA – genomic DNA

HNO – high nucleosome occupancy

IOD – inter-origin distance

LNO – low nucleosome occupancy

MFA – marker frequency analysis

OK-seq – Okazaki fragment sequencing

ORC – origin recognition complex

PCF – procyclic form

pre-RC – pre-replication complex

RSS – DNA replication start site

SNS – short nascent strands

SNS-seq – short nascent strands sequencing.

T4 PNK – T4 polynucleotide kinase

INTRODUCTION

DNA replication is one of the most fundamental processes by which a cell makes a complete copy of the genome before cell division. In eukaryotic cells, DNA replication initiates at multiple genomic sites, which are known as origins. Despite extensive research, a comprehensive universal code that is common for all eukaryotic origins remains elusive (reviewed in (1-4). The first isolated and characterized eukaryotic origin comes from budding yeast, *Saccharomyces cerevisiae* (5-7), where DNA replication starts from the AT-rich autonomously replicating sequence (ARS) with a loosely defined and thymidine-rich ARS consensus sequences (ACS) (8-10). The other eukaryotic origins lack a unique consensus sequence. Furthermore, studies on the metazoan origins have yielded disparate or discrepant results. Some studies reported increased AT-content (11-17) or long asymmetrical poly(dA:dT) tracts as replication initiation sites (18,19). Other reports describe the origins as regions with a high GC content, such as CpG islands (20-23) or regions near G-rich sequences with the potential to form G-quadruplexes (24-32). In addition, several epigenetic factors and chromatin environments were linked to eukaryotic origins, but without a common signature (1,2,33). However, replication origins exhibit several characteristics that distinguish them from the surrounding chromatin. They appear in nucleosome-depleted regions (34-37), which are flanked on one or both sides by well-positioned nucleosomes (18,35-38).

An active origin is characterized by the generation of two divergent single-stranded DNA fibres with short RNA primer at 5' end, known as short nascent strands (SNS) (reviewed in (39)). Size-selected SNS can be enriched and purified from broken chromosomal fragments by treatment with T4 polynucleotide kinase (T4 PNK) and λ-exonuclease, and this approach was developed for mapping of DNA replication initiation sites in yeast (40,41). Thereafter, this technique has been successfully applied in different species, using approaches such as SNS-on-chip (21,22) or by high throughput sequencing of SNS (SNS-seq) (23,24,26,28,31,32,42-45). A recent development has enhanced the conventional SNS-seq method by permitting the preservation of the directionality of SNS fibres following sequencing (23,43), and this improvement has augmented the potential of this approach to provide a more specific mapping of origins.

Trypanosoma brucei is a unicellular parasite of medical and veterinary importance that provokes deadly vector-borne diseases: sleeping sickness in humans, and nagana disease in livestock. This parasite circulates between mammalian and insect hosts, developing various life cycle stages: bloodstream form (BSF) in the mammals and the procyclic form (PSF) in the midgut of the tsetse fly (46,47). The order *Trypanosomatida* diverged from the main eukaryotic lineage about 600 million years ago (48,49) and displays some unusual genomic and epigenetic properties that provide valuable insight into the early processes employed by eukaryotic ancestors. Therefore, a better understanding of fundamental processes in divergent eukaryotes could not only reconstruct the early events of eukaryotic evolution (50,51) but also discern conserved mechanisms that are pivotal in higher eukaryotes. This could ultimately pave the way for the development of innovative medical treatments for sleeping sickness and nagana disease. The replication of DNA in this parasite has recently become a subject of interest and

our research group conducted a comparative analysis of DNA replication in some Trypanosomatids with mammalian cells using DNA combing (52). In a separate study, the initial attempt to identify the origins was conducted using marker frequency analysis coupled with deep sequencing (MFA-seq) (53-55), a technique designed to detect regions with increased DNA copy numbers. However, only a limited number of peaks were identified by MFA-seq. Therefore, a more sensitive approach is required to determine the position of DNA replication origins with higher resolution.

To map and identify the active origins of DNA replication (hereafter referred to as "origins") in *T. brucei*, we used a method that involves the isolation of RNA-primed short nascent strands (SNS) followed by sequencing that preserves the orientation of the SNS fibres. By employing this stranded SNS-seq methodology, we improved the resolution and the specificity of the origin mapping. We conducted an extensive bioinformatics analysis of mapped origins and compared our findings with previously published scientific datasets, which enabled us to identify several key components of DNA replication origins in *T. brucei* and to propose a model.

MATERIALS AND METHODS

Cell culture

Procyclic forms (PCF) of *T. brucei*, strain Lister 427 wild type strain were grown at 27 °C in SDM-79 (PAA Laboratories) supplemented with heat inactivated 10% FBS (Gibco) and 7 $\mu g.mL^{-1}$ hemin (56). Bloodstream forms (BSF) of *T. brucei* Lister 427 strain were cultivated at 37°C, 5% CO in a modified Iscove's culture medium (Gibco) complemented according to Hirumi *et al.* (57), with 10% (v/v) heat-inactivated foetal bovine serum (Gibco), 36 mM sodium bicarbonate (Sigma), 136 $\mu g.mL^{-1}$ hypoxanthine (Sigma), 39 $\mu g.mL^{-1}$ thymidine (Sigma), 110 $\mu g.mL^{-1}$ Na-pyruvate (Sigma), 28 $\mu g.mL^{-1}$ bathocuproine (Sigma), 0.25 mM β-mercaptoethanol, 2 mM L-cysteine.

SNS isolation and purification

The SNS were isolated and purified according to previously published protocol (58) with a few modifications. Between 2 - 2.5x10^9 cells were used as starting material for SNS isolation. Total DNA was isolated from three biological replicates of asynchronous PCF and BSF cells in the exponential phase of growth (3-5x10^6 and 7-9x10^5 $cells.mL^{-1}$ for PCF and BSF respectively). Cells were precipitated and incubated in 0.5 mL of 20 mM Tris-Cl, pH 7.0, 50 mM EDTA, 1% N-Lauroylsarcosine, 0.5% TRITON X-100 and 2 $mg.mL^{-1}$ proteinase K and incubated at 45°C for 24 hours. Fresh solution of proteinase K was added after 24 hours and digested for 16 hours more. DNA was purified with phenol/chloroform/isoamyl alcohol and precipitated in 0.3 M Na-acetate (pH 5.3) and absolute ethanol. DNA was dissolved gently in TEN20 buffer with 100 U RNasin (Promega). DNA was heat-denatured (95°C/5 min) and loaded onto 5 mL of neutral 5%–20% sucrose gradients prepared in TEN20 buffer in ultra-clear tubes (Beckman Coulter). Size separation of nucleic acids was done by ultra-centrifugation in SW50 rotor for 7 hours at 45 krpm at +4oC. Fractions of 200 µL from the sucrose gradient were collected and checked by agarose gel electrophoresis to assess the quality of nucleic acid separation

(Supplementary Figure 1A). Fractions from 0.5-2.5 x10^3 nt were pooled together as SNS enriched fractions. Half of the SNS pool was treated with 10 µg of cocktail of RNase, DNase free (Roche) for 16h at 37°C, to become the SNS-depleted control. SNS-enriched samples, together with its SNS-depleted controls, were treated 3-4 times with 50 U of T4 PNK (Thermo Fisher Scientific) and 100 U of λ-exonuclease (Thermo Fisher Scientific). Each treatment was done after heat denaturation of DNA and simultaneously on samples and controls. Finally, the rest of the RNA was removed with 5 µg of RNase, DNase free (Roche), SNS were sonicated with Bioruptor Pico sonication device (Diagenode) and purified with magnetic beads (CleanNGS, Proteigen). The quantity of purified SNS material was estimated by fluorometer DS 11 FX (DeNovix) using Qubit ssDNA Assay kit (Invitrogen).

It is important to note that we used SNS enrichment conditions and enzymatic treatments as described in previous reports (26,58). The conditions used here differ significantly from those applied in the reference (42). Here, we enriched the SNS by size on a sucrose gradient and then treated this SNS-enriched fraction with high amounts of λ-exonuclease. A previous study has shown that under these conditions, complete digestion of DNA containing G4-rich sequences is achieved (26). The efficiency of the T4 PNK and λ-exonuclease enzymes was tested on both sheared gDNA and on RNA. The majority of DNA was digested after one treatment with these two enzymes (Supplementary Figure 1B). RNA fibres treated under the same conditions were preserved (Supplementary Figure 1C), suggesting that SNS fibres would also be preserved. The quantity of DNA fibres in SNS containing fractions was systematically followed by qPCR after each enzymatic treatment and it was estimated that the amount of DNA in starting material was reduced around 100 times after the first treatment, and around 10 times more after each subsequent treatment with these two enzymes.

Preparation of stranded libraries for paired end sequencing

The NGS libraries from single stranded SNS were prepared according to the manufacturers protocol using the Accel-NGS 1S Plus DNA Library Kit and 1S Plus Dual Indexing Kit for Illumina (Swift Biosciences). A 3'end specific tail is preserving the strand orientation of the ssDNA fragments during library preparation, resulting in a second strand library, where the plus-strand fragments are represented by F1R2 read pairs and the minus-strand fragment by F2R1 read pairs. The quality of the libraries was checked by capillary electrophoresis on Bioanalyzer 2100 (Agilent) using Agilent High Sensitivity DNA Kit (Agilent Technologies). The quantity of libraries was estimated by fluorometry using dsDNA High Sensitivity Assay Kit on a DS-11 FX+ (DeNovix).

Paired end sequencing (2x150)

The libraries were validated for quality and quantity, using the Fragment Analyzer and qPCR techniques. qPCR was performed using KAPA Library Quantification Kits - Complete kit (Universal) (Roche, ref.7960140001) and Light Cycler 480 machine (Roche). The libraries were pooled equimolarly and sequenced on the Illumina NovaSeq 6000 using a NovaSeq Reagent Kit for 300 cycles (Illumina) following manufacturer's instructions. Libraries were sequenced on

one lane of a SP flow-cell in paired-end 150 nt. Sequencing was performed by the GenomiX (MGX) Core Facility, Montpellier, France. Demultiplexing and FASTQ files production was done using the Illumina software bcl2fastq (v2.20.0.422).

SNS-seq data analysis

Quality control of the FASTQ files was performed using FastQC (v0.11.9) (59). The sequencing reads were tail trimmed to remove the 3'end specific tail introduced during library preparation using cutadapt (60) with the following parameters (-u -10 -u 10 -U -10 -U 10). Afterwards the SNS-seq sequencing reads were aligned against the *T. brucei* Lister genome 427-2018 release 55 from TriTrypDB (https://tritrypdb.org/common/downloads/release-55/TbruceiLister427_2018, last accessed November 18th, 2024) with bowtie2 (v2.4.10) (61). The alignment output bam files were converted, sorted, and indexed using samtools (62). Duplicated reads were removed using picard (v2.23.5, "Picard Toolkit." 2019. Broad Institute, GitHub Repository. https://broadinstitute.github.io/picard/, last accessed November 18th, 2024; Broad Institute). Mapped deduplicated reads pairs were then separated with samtools *view* into minus-strand and plus-strand read pairs via their read orientation (minus-strand: F2R1 and plus-strand: F1R2). Peak calling was then performed with macs2 (v.2.2.7.1) (using default parameters plus -p 5e-2 -s 130 -m 10 30 --gsize 2.5e7) for plus and minus-strand reads separately.

Origin selection

Origins were selected after the peak calling step, in the second - peak filtering step. The called peaks were filtered using the bedtools suite (v.2.30.0) (63) as described below. In this peak filtering step, three criteria were defined to select the origins. The first criterion was that two SNS peaks must be oriented in a way that reflects the topology of the SNS strands within an active origin: the upstream peak must be on the minus strand and followed by the downstream peak on the plus strand. The second criterion was to set a maximal inner to inner border distance between to divergent minus and plus peak pair. The distances between the two divergent leading strands are not known. We have calculated that the average median inner to inner distance between minus and plus peaks in the replicates was 0.68kb. We conducted tests with several different inner border to border distances ranging from 0.5 kb to 2.5 kb. Finally, we applied 0.5 kb inner border to border distance because many of the neighbouring peaks were in that window. The third criterion was introduced to prevent complete overlap of two divergent peaks, as this can only occur if the SNS fibres elongate at the same rate in both directions, which is an unlikely scenario. Indeed, the synthesis of the Okazaki fragment and ligation to the 5' of the SNS requires several enzymatic processes, and we can assume that the elongation of SNS fibre is slower at the 5' end. Of note, complete overlap of minus and plus peaks was rarely observed (mean 6.5% (stdev 6.5%, min 0.8%, max 15.8%)). To define the maximum overlap between the minus and plus SNS peaks, a series of tests were conducted, varying the percentage of overlap from 40% to 90%. This resulted in a similar number of origins, and ultimately, 50% was selected as the optimal value. Therefore, firstly, minus-strand and plus-strand peaks with an overlap of over 50% were removed with bedtools *intersect*. Secondly peak

pairs on different strands with a maximal distance of 0.5 kb were selected using bedtools *windows*. Finally, peak pairs were filtered for the correct orientation (minus-strand followed by plus-strand) with bedtools *closest* and the origin was defined as the genomic region between the two inner borders of two divergent peaks. The overlap of origins between replicates and cell types as well as other genomic features was analysed with bedtools *windows* and bedtools *intersect*. Replicates were merged with bedtools *merge*.

For **visual inspection of genomic features** IVG (64) was used. Violin plots were generated with the GraphPad Prism 8.3.0 software (GraphPad Software Inc., La Jolla, CA, USA).

Shuffled controls were generated with bedtools *shuffle* (-chrom –noOverlapping –seed -excl) that randomly permutes genomic features provided respecting the number, the size and chromosomal location and either excluding the detected origins or excluding detected origins and coding sequences (called intergenic shuffled control).

Clustering of origins origin clustering was analysed with bedtools *cluster* using different distances.

The inter-origin distance (IOD) between the origins detected by SNS-seq was calculated using bedtools *closest* (-io (ignore overlapping) -iu (ignore upstream)) on a single file.

PCF down-sampling was performed to be able to compare origin number in BSF and PCF. The mapped reads from all three PCF and BSF replicates, respectively, were merged using Samtools *merge*. The PCF combined sample was sub-samples with Samtools *view* (--subsample 0.69 --subsample-seed [0-10]) 15 times to obtain a set of mapped reads comparable to the BSF combined sample. Origin detection was applied as described above to the combined BSF and the 15 subsamples combined PCF samples.

For **visualization of the respective** localization of two or more datasets the the Deeptools suite (65) and the ggplot2 R package (66) was used. In Deeptools bed and bigwig files were used to generate a matrix file with *ComputeMatrix* and *PlotHeatmap* or *PlotProfil* were used to generate visualizations from the matrix. The open source ggplot2 R package (66), freely available from the Comprehensive R Archive Network (CRAN) repository (http://www.r-project.org, last accessed November 18th, 2024) was used to generate composite plots.

Motive enrichments were tested over the origin centres extended 250 bp up and downstream with the MEME suite motive discovery (67) as well as with homer *findMotifs* (68). A multiple sequence alignment and the analysis of nucleotide patterns were built with the WebLogo online application (69).

Nucleotide composition around the origin centres was calculated as follows. The *T brucei* genome was split into 20 bp windows using bedtools *window*. The number of A, T, G and C nucleotides for each window were obtained with bedtools *nuc* and used to calculate the nucleotide percentage per 20 bp window. The resulting bedgraph files were transformed into bigwig file format using USCS utilities *bedGraphToBigWig*

(https://github.com/ucscGenomeBrowser/kent/tree/master/src/utils, last accessed November 18th, 2024).

Comparison of origins mapped by stranded SNS-seq with previously published datasets.
MNase-seq. The MNase-seq data was downloaded in as bigwig files from GEO (accession number GSE90593) (70). The data was analysed using the *T. brucei* 427 reference genome. To be able to compare our SNS-seq dataset with this data we reanalysed our sequencing reads as described above in the same genetic background (https://tritrypdb.org/common/downloads/release-62/TbruceiLister427/, last accessed November 18th, 2024). To quantify the MNase-seq signature we observed around the origin centres ("M" pattern), we calculated the signal foldchange between the regions 400-200nt upstream and 200-400 nt downstream of the origin centre with the signal 75nt up- and downstream around the centre of the origins. A fold change of 1.5 was considered positive. The accumulated signal values for these regions were extracted from the MNase-seq bigwig files with the UCSC utilities *bigWigAverageOverBed* command (https://github.com/ucscGenomeBrowser/kent/tree/master/src/utils, last accessed November 18th, 2024).

MFA. Two MFA datasets were obtained, one was obtained directly from the Authors (53), and one was downloaded from ENA (accession number PRJEB11437) (54), both in form of fastq files. The analysis of the fastq files was performed as described previously (53,54). The original published MFA analysis was performed in *T. brucei* 927 genomic backgrounds. We re-analysed the data in the 927 (https://tritrypdb.org/common/downloads/release-62/TbruceiTREU927/, last accessed November 18th, 2024) and the 427-2018 genomic background.

Drip-seq. The Drip-seq dataset was downloaded from ENA (accession number PRJEB21868) (71) as fastq files. The re-analysis was done as described (71). The original published Drip-seq analysis was performed in *T. brucei* 427 genomic backgrounds. We re-analysed the data in the 427 and the 427-2018 genomic background.

G4-seq. The G4-seq (72) dataset was downloaded from GEO (accession number GSE110582) as bed and bedgraph files separated for the plus and minus strand.

G4 prediction in the *T. brucei* genome were done with G4Hunter (73) with the following settings, window size (-w 25) and a threshold of 1.57.

DNA molecular combing

Asynchronous cell populations of *T. brucei* were grown to the density of $4x10^6$ and $7x10^5$ cells.mL^{-1} for PCF and BSF, respectively. Cells were sequentially labelled with two modified nucleosides, firstly with 300 µM iodo-deoxyuridine (IdU, Sigma) for 20 minutes and then 20 minutes with 300 µM chloro-deoxyuridine (CldU, Sigma) without intermediate washing. After labelling, the cells were immediately placed on ice to stop DNA replication and extensively washed with ice-cold 1xPBS. Around $1x10^8$ cells were blocked in 1% low melting agarose plugs.

DNA molecular combing was done as described previously (52). Protein-free plugs with DNA were extensively washed in TEN20 buffer (10 mM Tris-Cl, 50 mM EDTA, 20 mM NaCl, pH 7.0) and then stained with 0.5 µM YOYO-1 fluorescent dye (Molecular Probes) for 1 h at 37°C. After washing from the excess YOYO-1 day, the 200 µL of TEN20 buffer was added to the plugs, melted at 65 °C for 15 minutes, slowly cooled down to 42 °C and treated with 10 U of β-agarase (Merck) for 16h. This treatment was repeated with fresh β-agarase for 4 hours. After digestion, 2 mL of 50 mM MES (mix of 0.5 M MES Hydrate and 0.5M MES Sodium Salt, pH 5.7) was added very gently to the DNA solution and then DNA fibres were combed and regularly stretched (2 kb.µm^{-1}) on silanized coverslips provided by Montpellier DNA Combing Facility. Linearized DNA fibres were fixed for 16 hours at 65 °C, denatured in 1N NaOH for 20 minutes and blocked with 1 × PBS/ 1% BSA/ 0.1% Triton X100. Immuno-detection was done with antibodies diluted in 1 × PBS/ 1% BSA/ 0.1% Triton X100 and incubated at 37 °C in a humid chamber for 60 min. Each step of incubation with antibodies was followed by extensive washes with 1 × PBS. Immuno-detection was done with anti-ssDNA antibody (1/100 dilution, clone 16-19, Merck Milipore), the mouse anti-BrdU antibody (1/20 dilution, clone B44, Becton Dickinson) and rat anti-BrdU antibody (1/20 dilution, clone BU1/75, Eurobio Scientific). The following secondary antibodies were used: Alexa488 Goat anti Rat IgG (1/50 dilution, Molecular Probes), Alexa 546 Goat anti Mouse IgG1 (1/50 dilution, Molecular Probes) and Alexa 647 goat anti-mouse IgG2a (1/100 dilution, Molecular Probes). Coverslips were mounted with 15 µL of Prolong Gold Antifade (Molecular Probes) and dried at room temperature for 16 hrs. Images were acquired using Zeiss Z1 microscope, equipped with an ORCA-Flash 4.0 digital CMOS camera (Hamamatsu), using a 40× objective, where 1 pixel corresponds to 325 bp (one microscope field of view corresponds to ~450 kb). Acquisition and analysis of images were done by MetaMorph version 7 software (Molecular Devices LLC).

Measurements and statistical analysis of combed fibres

The speed of replication forks was estimated on individual forks, presented as IdU track followed by a CldU track. Only unbroken forks were measured, which was followed by DNA immuno-staining. The speed was calculated as the ratio between the CldU track length (kilobases) and the CldU pulse duration (minutes). The IODs were measured as the distance (in kilobases) between the centres of two adjacent progressing forks located on an uninterrupted DNA fibre. Fork asymmetry was calculated as the ratio of the longer track over the shorter track in progressing divergent forks. A ratio greater than one indicates asymmetry. The GraphPad Prism 8.3.0 software (GraphPad Software Inc., La Jolla, CA, USA) was used to generate graphs and to perform statistical analysis of DNA combing measurements as described previously (52). The statistical comparisons of different datasets were assessed by the non-parametric Mann–Whitney two-tailed tests.

RESULTS

Origin mapping by stranded SNS-seq.

The stranded SNS-seq method was employed to map the origins in two distinct cell types of *T. brucei*, representing two life cycle stages of this parasite: PCF and BSF cells. The experimental workflow for purifying and enriching SNS, together with the corresponding control, is illustrated in Figure 1A. An SNS-enriched sample was prepared through a two-stage process. Firstly, the sample was enriched by size, and secondly, any broken DNA fibres were removed through λ-exonuclease treatments. An SNS-depleted sample was generated as a background control, comprising fibres that were resistant to RNase and/or λ-exonuclease treatments (Materials and Methods and Figure 1A). After preparation of samples and controls, stranded libraries were prepared for high-throughput sequencing. Each single stranded DNA molecule was differentially tagged at 3′ and 5′ end to preserve the orientation of the DNA fragments after sequencing (Materials and Methods). Paired-end sequencing was then performed on the three biological replicates of PCF and BSF cells, with each replicate containing an SNS-enriched sample and SNS-depleted control. The starting number of reads after paired-end sequencing and the number of reads after each step of the bioinformatic analysis are shown for each sample in Supplementary Table 1 and Supplementary Figure 2A.

The bioinformatics workflow used here for origin selection is schematically presented in Figure 1B (see also Materials and Methods). In brief, after the alignment and sorting of the reads to the plus and minus strands, two steps of peak selection were performed. The first step was strand-specific peak calling against SNS-depleted control to remove background noise. The second step, designated as "peak filtering", is employed to identify pairs of peaks that align with the divergent orientation of SNS fibres in an active origin (Figure 1C). In accordance with the criteria delineated in the peak filtering step (see Materials and Methods and Figure 1B), any peaks identified as significant in the peak calling step that were not in a divergently oriented pair were excluded from the analysis as false positives. The proportion of removed called peaks after peak filtering ranged from 88 to 96% (Supplementary Figure 2B and 2C). The peak filtering step utilised the information about the divergent orientation of SNS fibres, a key characteristic of active origins, in origin mapping, to enhance the specificity of origin mapping. Ultimately, the origins were identified as genomic regions situated between the inner borders of two closely positioned and oppositely oriented peaks (Figure 1C).

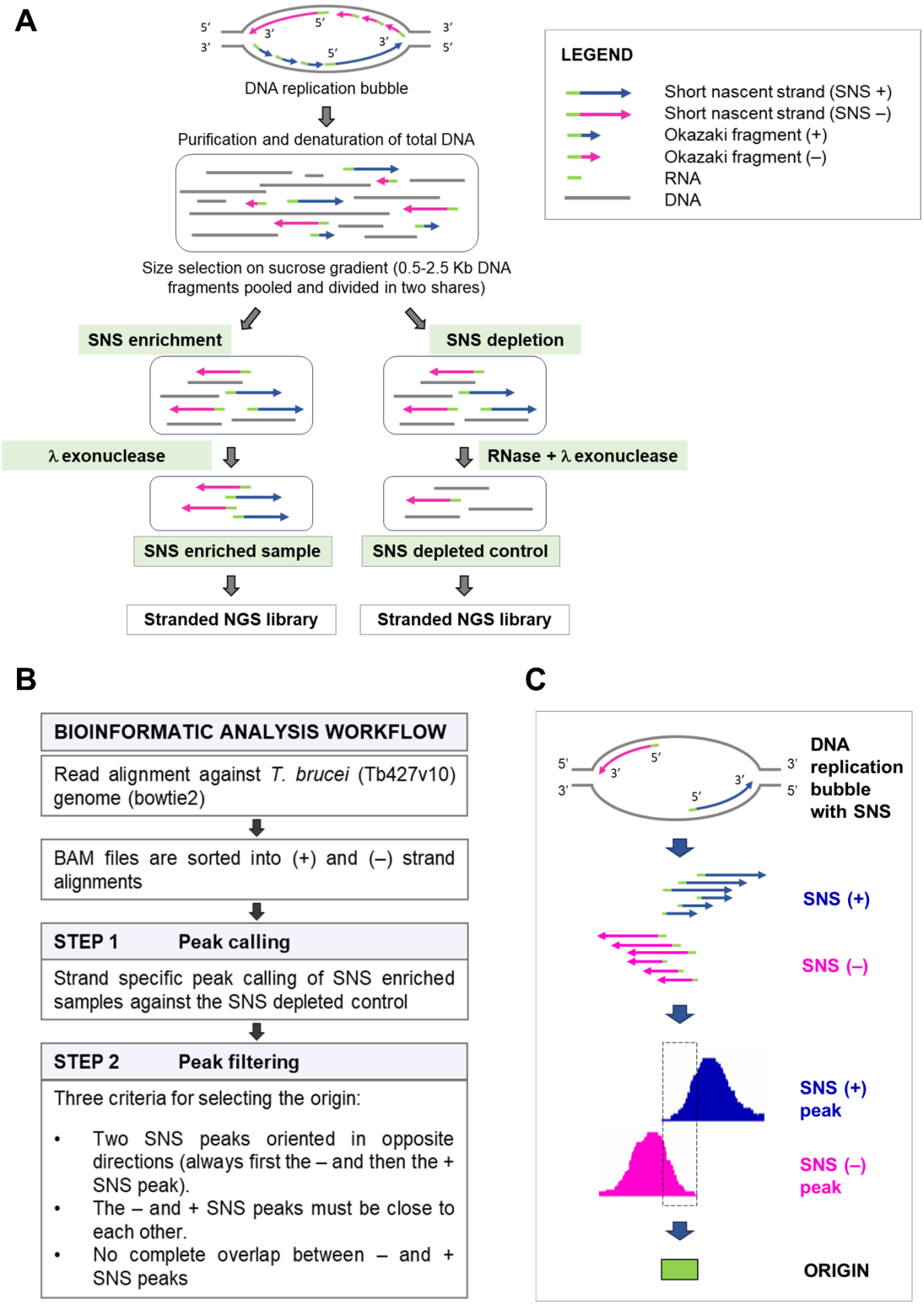


Figure 1. Experimental protocol and bioinformatics workflow for selecting the origins.
A) Schematic representation of the experimental workflow for short nascent strand (SNS) purification. A replication bubble with possible DNA replication intermediates and their orientation is shown at the top.

The methodology employed for the generation of the SNS-enriched samples and corresponding SNS-depleted controls is illustrated. B) Schematic diagram of the bioinformatic workflow applied for stranded SNS-seq data analysis. Two steps of peak selection were performed to map the origins. The first step was peak calling against SNS depleted control. The second step was peak filtering, based on the three indicated criteria. C) The schematic representation shows one origin, generating a replication bubble with two SNS leading fibres in divergent orientation (SNS + and SNS –). The plus and minus SNS fibres are presented to demonstrate the growth of the leading strand in both directions after the ligation of the upstream Okazaki fragments to the 5′ ends of SNS. The different sizes of the isolated SNS fibres are also presented. The pink peak represents the SNS minus peak synthesized on the plus DNA strand. The blue peak represents the SNS plus peak synthesized on the minus DNA strand. The dashed rectangle, situated between the two inner borders of two divergent SNS peaks, and the green box represent the mapped origin.

Figure 2A presents a representative screenshot of the distribution of peaks and positions of the mapped origins for three replicates of BSF cells. The distribution of the lengths of the mapped origins was estimated for each replicate (Supplementary Figure 2D), and the average length of origins was found to be approximately 150 bp, representing a substantial improvement in resolution over previous origin mapping studies. Following the mapping, the origins from the three replicates were merged, resulting in the identification of a total of 914 origins in BSF cells and 549 in PCF cells. Our next objective was to identify the preferred genomic sites of origin activation. To this end, the distribution of origins was analysed around centred intergenic and genic regions. The results showed that the origins were unevenly distributed throughout the genome, with a higher density in intergenic regions and a lower density in genic regions (Figure 2B). The same enrichment of origins in intergenic regions was observed in both cell types and was absent in the shuffled controls, which presented random genomic regions chosen with respect to the size, number and chromosomal distribution of origins (Figure 2B). Subsequently, the proportion of origins in four distinct genomic regions was determined, revealing that over 92% of PCF and 85% of BSF origins map within intergenic regions. The minority were found in genes and untranslated regions (UTRs) of genes (5.8% for PCF and 9% for BSF), pseudogenes (1.5% for PCF and 5.9% for BSF) and noncoding RNA regions (0.2% for PCF and 0% for BSF) (Figure 2C). The observed distribution of origins was found to be significantly different from the distribution that would be expected if the origins were distributed evenly across the genomic sequence (P-value<0.0001) (Figure 2C). Subsequently, the clustering of origins was examined within genomic regions of increasing length. The findings showed that the origins in *T. brucei* were organised into clusters (Supplementary Figure 2E) and that origin clustering was significantly different from the clustering of shuffled controls (Figure 2D). The results showed that 50% of the PCF origins were clustered within 11.5 kb regions (shuffled control median was 17 kb), while 50% of the BSF origins were clustered within 6.5 kb regions (shuffled control median was 13 kb) (Figure 2D). This finding was in line with previous research in other organisms, which showed that origins are unevenly distributed and clustered (26,28,31). Next, the intersection of the two sets of origins revealed the proportion of common origins between the two cell types, showing that 45% of PCF origins were also active in BSF cells (Supplementary Figure 3A). The number of origins in the two cells was compared using normalized mapped reads (Materials and Methods). This normalization

showed that BSF cells have 45% more active origins than PCF cells (Supplementary Figure 3B). Afterwards, the IOD distribution at the population level was calculated for both cell types. The calculated population-based IODs were shorter in BSF (median 15.3 kb) compared to PCF cells (median 21.4 kb) and this difference was statistically significant (Supplementary Figure 3C). However, very long IODs were observed, albeit at low frequencies, reaching up to 354 kb in PCF and 488 kb in BSF (Supplementary Figure 3C).

Despite considerable discrepancies in the origin number and resolution, a comparison was conducted between the origins identified by stranded SNS-seq and those previously detected by MFA-seq. In *T. brucei*, MFA-seq identified 42 broad peaks (53-55), which is insufficient for replicating the entire genome in a timely manner. Upon intersection of the two datasets (Supplementary Figure 4), it was calculated that 29-53% of origins detected by stranded SNS-seq were situated within broad MFA peaks. The percentage varied depending on the MFA dataset analysed. However, the observed overlap was not statistically significant when compared to the shuffled control (P-values > 0.4, two-tailed t-test).

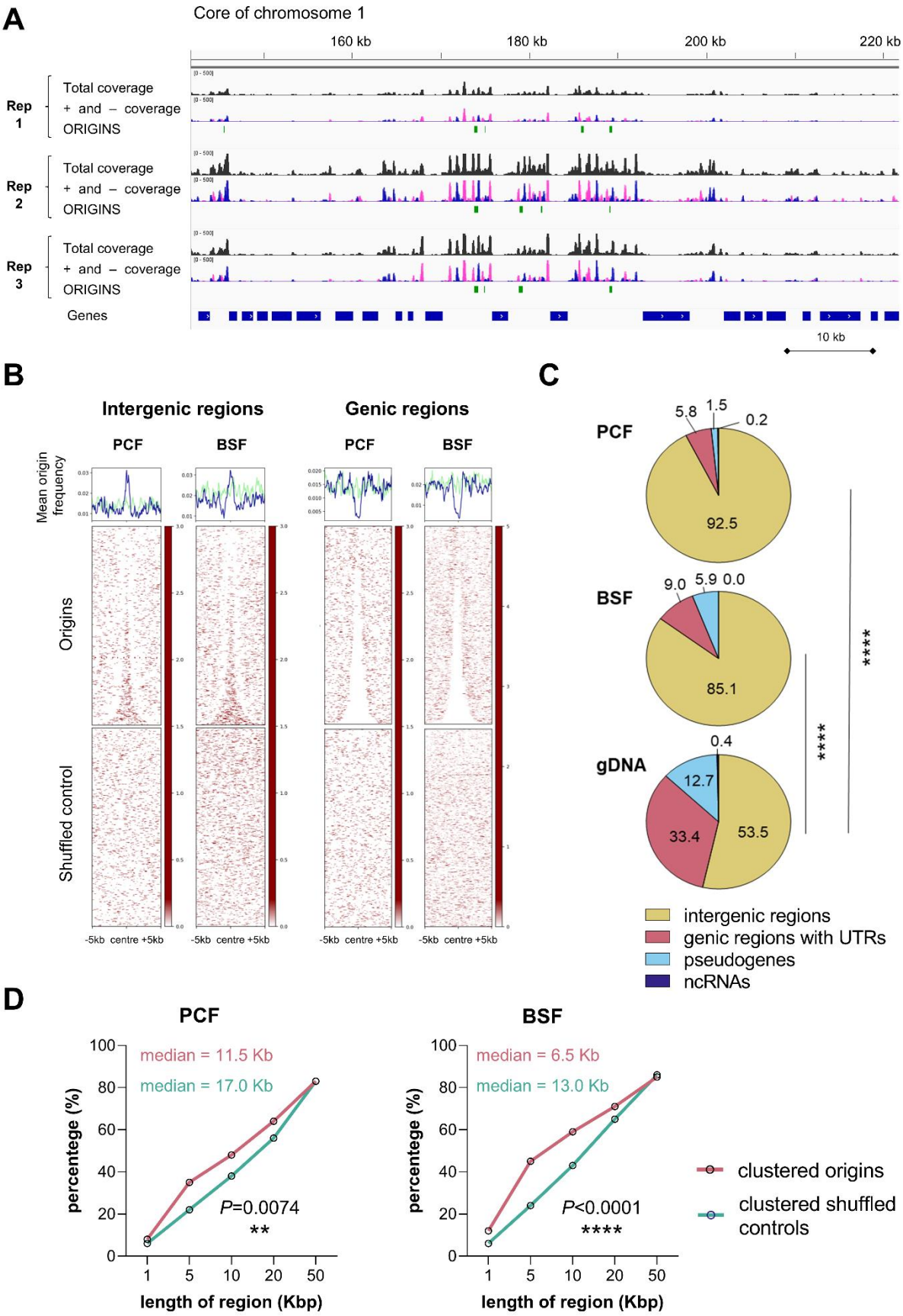


Figure 2. Genomic distribution of mapped origins.

A) A representative screenshot of the read coverage and the positions of the mapped origins in a ~80 kb long genomic region of the core chromosome 1. The three replicates of BSF cells are presented. The total read coverage is in black, the minus and plus strand read coverage are in pink and blue, respectively. The green bars represent the mapped origins. B) Heatmaps present the distribution of the mapped origins and shuffled controls within centred genic and intergenic regions (± 5 kb). Shuffled controls present random genomic regions chosen with respect to the size and chromosomal distribution of origins. C) The pie charts show the proportions of origins within four indicated genomic regions. The gDNA presents the proportions of four genomic regions within genomic sequence. The *P*-values were computed using Chi-square tests, which involved a comparison of the absolute numbers of indicated categories for each pair of datasets (**** - $P<0.0001$). D) The distribution of clustered origins and shuffled controls within genomic regions of increasing lengths. The *P*-values were computed using Chi-square tests.

Insights from DNA molecular combing: comparative analysis of single cell and cell population origin usage.

Stranded SNS-seq and DNA molecular combing are two complementary techniques. Stranded SNS-seq provides a genome-wide overview of active origins within a given cell population. However, it should be noted that the usage of these origins may vary between individual cells within the same population. In contrast to the first technique, DNA combing provides an overview of DNA replication parameters at the single cell level. In this study, we used DNA combing to compare the replication of PCF and BSF cells. Replicating cells were labelled with two modified nucleosides, followed by immuno-detection (Materials and Methods). After immuno-detection, the three discrete signals were discerned on combed DNA fibres. The red signal, followed by green, indicated the direction of replication forks, while the blue signal represented the unmodified DNA fibres (Figure 3A). The origins were identified at the centre of two divergently oriented replication forks. The DNA replication parameters, including the speed of replication forks, IOD and asymmetry of replication forks, were determined as previously described (52). The median replication fork velocity was 1.73 and 2.44 $kb.min^{-1}$ for PCF and BSF, respectively, while the median IOD was 152 and 213 kb for PCF and BSF, respectively (Supplementary Table 2). The median speed of replication forks was 1.4 times higher, and the median IOD was 1.4 times longer in BSF compared to PCF cells, indicating statistically significant differences between the two cell types (Figure 3B and Figure 3C). Moreover, we compared the frequency of asymmetric replication forks, defined as bidirectional forks that progress from the centre of origin at unequal rates, between the PCF and BSF cell types and found no statistically significant differences (Figure 3D). Given the high sensitivity of DNA combing measurements to the DNA fibre lengths (74), a comparison of the analysed DNA fibre lengths was performed as a control. This analysis showed that there was no statistically significant difference in the length of the analysed fibres between the two cell types (Figure 3E, Supplementary Table 2). This control confirmed that the observed differences in velocity and IODs were not an artefact of the different lengths of the analysed fibres, but rather reflected genuine differences between the replication of PCF and BSF cells.

The approximate number of active origins in a single cell can be estimated by dividing the length of genomic sequences (~50 Mb) by the mean or median IODs obtained by DNA combing. It was thus estimated that the range of active origins in a single cell was 279-329 for

PCF and 228-235 for BSF cells (Supplementary Table 2). Once the number of origins used by individual cells had been estimated, our subsequent objective was to compare these figures with the numbers of mapped origins obtained by stranded SNS-seq at the cell-population level. The stranded SNS-seq method revealed 549 origins for PCF and 914 origins for BSF, representing a two- to four-fold increase in active origins in the cell-population compared to a single-cell level. Therefore, our estimate shows that a single PCF cell activated 51-59%, while a BSF cell activated only 25-26% of the origins that were used at the entire cell-population level. This phenomenon, which is known as flexible origin usage, is a well-documented characteristic of eukaryotic origins (reviewed in (75)). The higher origin flexibility of BSF cells can be attributed to the shorter cell cycle and consequently shorter S phase, during which the genome must be duplicated.

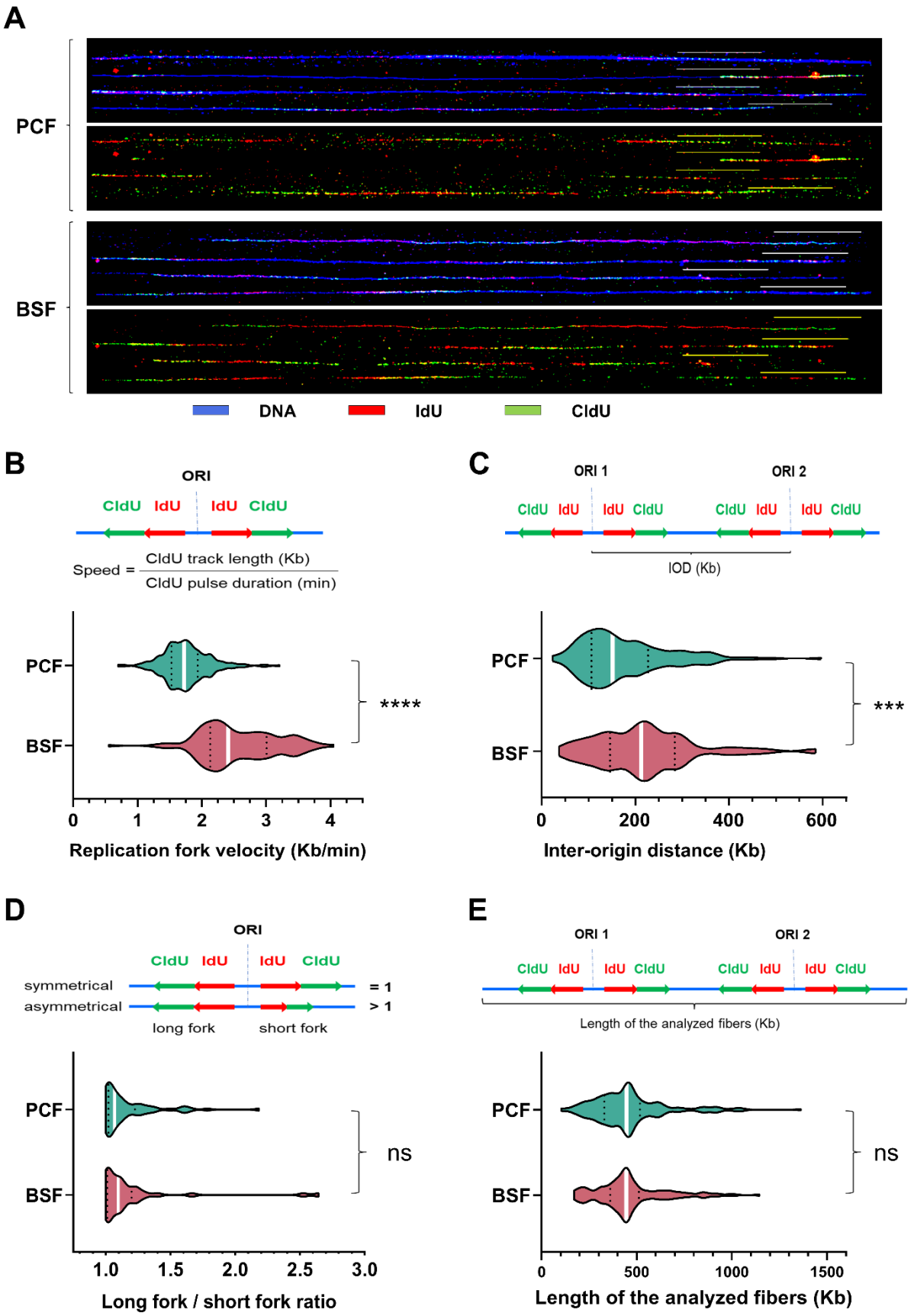


Figure 3. Comparison of DNA replication parameters between PCF and BSF cells by DNA molecular combing.

A) The figure depicts representative chromosome fibres. The immuno-detected DNA fibres are indicated in blue, the initial pulse of nucleosides (IdU) in red, and the subsequent pulse of nucleosides (CldU) in green. The lower panels display only the tracks from the first and second pulse of nucleosides extracted from the corresponding upper panels. A scale bar of 50 kb is marked on each cropped DNA fibre. B) The velocity of the replication forks was calculated by dividing the length of the CldU tracks with the duration of the CldU pulse on unbroken fibres (schema on upper panel). Violin plots present distribution of the measured replication fork velocities in two cell types. C) The inter-origin distance (IOD) was defined as the length between two adjacent replication initiation sites and can be determined by measuring the centre-to-centre distances between two adjacent progressing forks (schema on upper panel). Violin plots present the distribution of the measured IODs in two cell types. D) The upper panel presents the concept of long fork/short fork ratios. The long fork/short fork ratio corresponds to the ratio of the longer IdU + CldU signal length over the shorter IdU + CldU signal length of bidirectional replication forks. A ratio >1 indicates fork asymmetry, while a ratio =1 indicates fork symmetry (52). The lower panel presents the distribution of the measured asymmetry of replication forks in two cell types. E) The upper panel presents the concept of the measurements of the lengths of the analysed fibres. The lower panel presents violin plots with the measured lengths of the analysed fibres in two cell types. White bars on the violin plots indicate median values. Dotted black lines indicate quartiles. Two-tailed Mann-Whitney test was used to compute the corresponding *P*-values (ns - non-significant; * - $P<0.05$; ** - $P<0.01$; *** - $P<0.001$; **** - $P<0.0001$).

DNA replication starts between poly(dA) and poly(dT) enriched sequences.

Once the origins of PCF and BSF cells had been mapped, our subsequent objective was to identify the common elements that would define the origins of *T. brucei*. To achieve this, the origins from both cells were merged into a single set, comprising a total of 1225 origins. As most origins were identified within intergenic regions, a novel shuffled control was developed to represent random intergenic regions selected in accordance with the number of origins, their size, and chromosomal location (Materials and Methods). The mean distance of origins and intergenic shuffled controls from the genes was found to be similar (745.2 bp for origins and 797.6 bp for shuffled controls), thus providing a more accurate control. Accordingly, all subsequent results will be presented using this merged set of origins and the intergenic shuffled controls. Our initial objective was to examine the nucleotide distribution of centred origins and to compare it with the shuffled controls. The resulting plots showed that the percentage of adenines (A) or thymines (T) was higher near the centre of the origins (Figure 4A). This percentage was well above 40%, which was higher than the mean A or T content of the intergenic regions (reaching 30%) (Figure 4A), while the overall A or T content of the genomic sequence is approximately 28% (76). The enrichment near origins showed a remarkable sequence polarity represented by a strong enrichment of A upstream and a strong enrichment of T downstream of the origin centre (Figure 4A, Supplementary Figure 5A). Furthermore, a remarkable polarity in the distribution of guanines (G) and cytosines (C) was also observed. The origins exhibited a slight enrichment of G nucleotides upstream (up to 24%, while the average intergenic G was 20%), coinciding with the A enrichment, and a slight enrichment of the C nucleotides (up to 24%, while the average intergenic C was 20%), coinciding with the T enrichment downstream of the centres (Figure 4A). The nucleotide pattern that was specific for origins was entirely absent in the shuffled controls (Figure 4A, Supplementary Figure 5A). The subsequent step was to search for consensus motives of origins (Materials and Methods). However, apart from extended sequences that were enriched with

adenines or thymines (data not shown), no discernible consensus motif was identified. The sequences surrounding the origins were also analysed using the WebLogo software (69), and a graphical representation of nucleotide occurrence unambiguously confirmed an unequal nucleotide distribution, with an enrichment of A and G upstream and T and C downstream of the centre of origins (Supplementary Figure 5B). After that, the distribution of polynucleotides of varying lengths was examined in the vicinity of the origins. The results demonstrated the presence of poly(dA) enriched sequences upstream and poly(dT) enriched sequences downstream of the centre of origins, which was not detected in the shuffled controls (Figure 4B, Supplementary Figure 6A). It is worthy of note that the poly(dA) enrichment and an enrichment of up to four Gs were present upstream, while poly(dT) and up to four Cs in a row were enriched downstream of the origin centres (Figure 4B). Accordingly, a single origin can be defined as a genomic site with a specific nucleotide skew, characterised by an overrepresentation of poly(dA)s and poly(dG)s upstream, and poly(dT)s and poly(dC)s downstream of the origin centre.

Subsequently, the proportions of origins that lacked or possessed four As upstream and/or four Ts downstream of the centre were calculated and compared with the intergenic shuffled controls (Figure 4C, left graph). The same calculation was performed for eight As and Ts (Figure 4C, right graph). The findings revealed that 1149 origins exhibited the presence of four As upstream and four Ts downstream of the centre, representing a proportion of ~94% compared to ~78% for the intergenic shuffled controls. Similarly, 450 origins (~37%) possessed eight As upstream and eight Ts downstream of the centre, compared to only 114 (~9%) intergenic shuffled controls with this pattern. Accordingly, the distribution of poly(dA) and poly(dT) observed for origins was found to be significantly different from that observed for the intergenic shuffled control (Figure 4C).

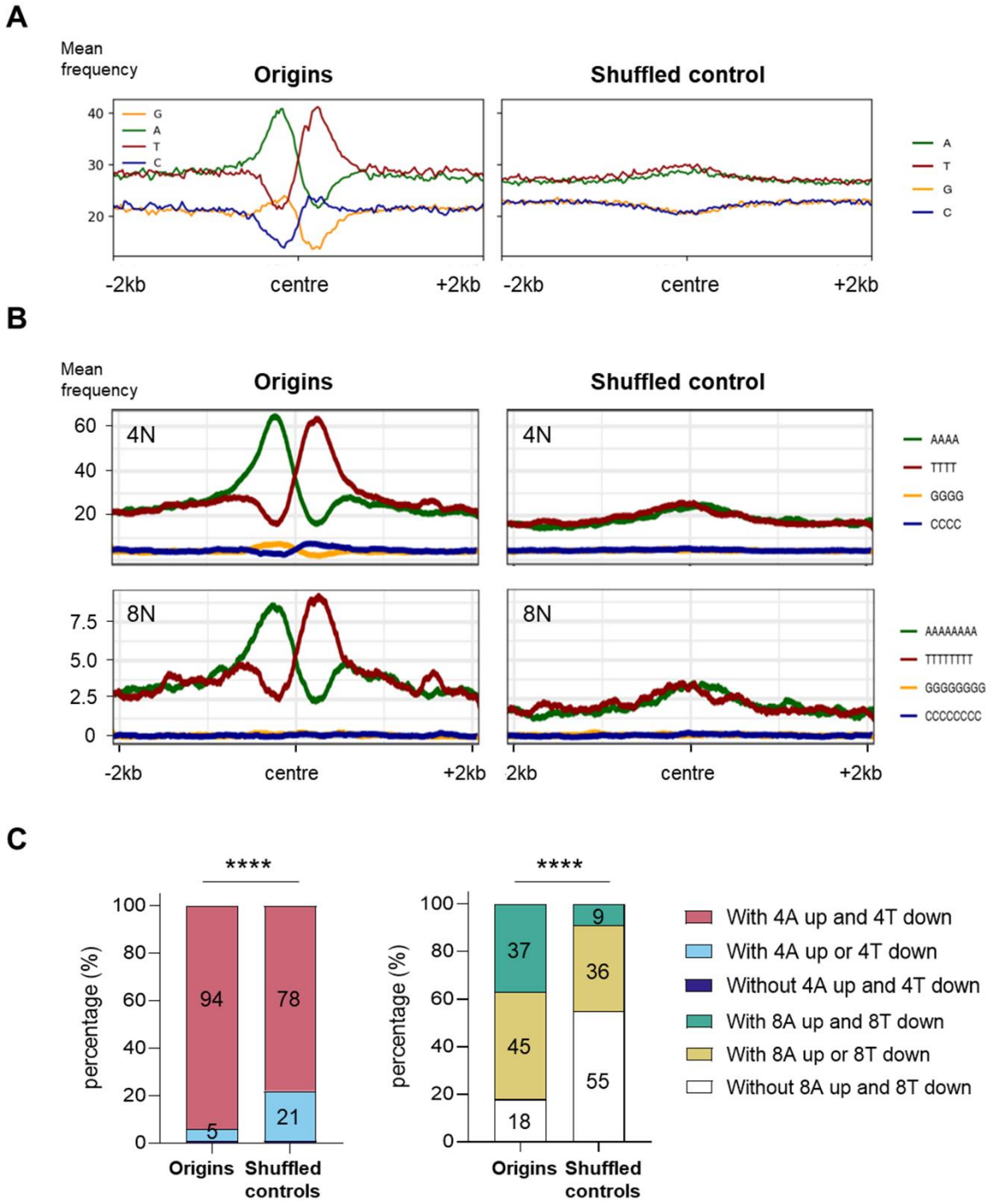


Figure 4. The spatial organisation of nucleotides and polynucleotides in the vicinity of origins.

A) The profile plots illustrate the distribution of four nucleotides around centred origins and shuffled controls (± 2 kb). The nucleotide percentage was calculated within 20 bp window B) The profile plots illustrate the distribution of four and eight polynucleotides around centred origins and shuffled controls (± 2 kb). A smoothing function was employed to calculate the mean frequency of polynucleotides per position within a 100 bp window (50 bp up- and downstream from the position). C) The proportions of origins and shuffled controls without or with four or eight As upstream and/or four or eight Ts downstream of the centre. The ±0.5 kb window from the centre was analysed. The *P*-values were

calculated using the Chi-square test, which involved a comparison of the absolute numbers of indicated categories for each pair of datasets (**** - $P<0.0001$).

G4 structures were accumulated in a specific pattern in proximity to the origin.

G-quadruplexes (G4s), DNA secondary structures with non-canonical Watson-Crick base pairing, have recently emerged as potential regulators of genome replication (77,78). Given the previously reported presence of G4 structures in the vicinity of some eukaryotic origins (24,25,27,32), an investigation was undertaken to determine the presence of G4s in the vicinity of the mapped origins in *T. brucei*. Our analysis employed two sets of previously determined G4s by G-quadruplex sequencing (G4-seq) (72). One set of G4s was obtained under physiological conditions, while the other was obtained in the presence of pyridostatin (PDS), a drug that binds and stabilises G4 structures (72). The origin dataset was intersected with the stranded G4s datasets obtained in both conditions, resulting in a notable enrichment of G4s in the vicinity of the centre of origins (Figure 5A). A distinctive pattern of G4 distribution was observed, whereby the G4s on the plus strand were found to be enriched upstream, while the G4s on the minus strand exhibited an enrichment downstream of the centres of origin (Figure 5A). This pattern of G4 enrichment was absent in the intergenic shuffled control. The occurrence of G4s in the vicinity of the origins exhibited variability, with a discernible increase in the number of G4s near centre observed under PDS-stabilised conditions in comparison to physiological conditions (Figure 5A).

Our subsequent objective was to determine the proportion of origins that possessed G4s on one or both sides and those that lacked G4s. To achieve this, the region of ±2 kb around the centres of origins was analysed and compared to shuffled controls. The result showed that the origins displayed significantly different G4 distributions compared to the shuffled controls ($P<0.0001$) (Figure 5B). Furthermore, the proportions were significantly different in PDS-stabilised compared to physiological conditions (Figure 5B). A set of origins with G4s on both sides constituted a minor population under physiological conditions (9%), whereas this was a major population in PDS-stabilised conditions (39%). The proportion of origins with G4 on one or both sides increased significantly in PDS-stabilised condition (74%), compared to the physiological condition (47%) (Figure 5B). Analyses of the presence of G4s in the region ±0.5 kb around the centres of origins yielded comparable results (Supplementary Figure 6B).

To gain a deeper insight into the distribution of origins and the G4 structures, an empirical cumulative distribution functions (ECDF) were computed. The ECDF demonstrated that the distances between the origins and the closest G4s were smaller than those observed for the shuffled controls (for physiological G4s condition - median of 2010 bp for origins versus median of 4154 bp for intergenic shuffled controls and for stabilised G4s condition – median of 329 bp for origins versus median of 1223 bp for intergenic shuffled controls) (Figure 5C). This implies that 50% of the origins possesses a stabilized G4 structure at ±329 bp from the centre. Conversely, 80% of origins possess a stabilised G4 structure at ±2575 bp from the centre (±9354 bp for the shuffled control). Accordingly, a strong association was identified between

origins and physiological G4 structures (odds ratio = 1.87; *P*-value = 0), as well as for stabilised G4s structures (odds ratio = 2.13; *P*-value = 0).

Our results demonstrated that poly(dA) enriched sequences and G4s were enriched upstream from the centre of origins on the plus strand, and downstream on the minus strand (Figure 4B and Figure 5A). Therefore, we wanted to determine the spatial relationship between these two elements. To achieve this, the strand separated poly(dA) enriched sequences and the experimental G4s were intersected at centred origins and shuffled controls. The results demonstrated that the peaks of the experimental G4s and poly(dA) sequences overlapped near the centres of origins in a strand specific manner that was notably absent in the intergenic shuffled controls (Figure 5D). Our findings revealed a previously unidentified element of origin: the enrichment of poly(dA) and G4 structures exhibited an important association upstream of the centre on the plus strand and downstream of the centre of origin on the minus strand (Figure 5D). It was observed that the poly(dA) peaks exhibited summits between ±100 to 300 nt, while the G4 peaks exhibited summits in the range of ±100 to 200 nt from the centres of origins (Figure 5D). Such regions were designated as poly(dA)/G4-enriched elements of origins.

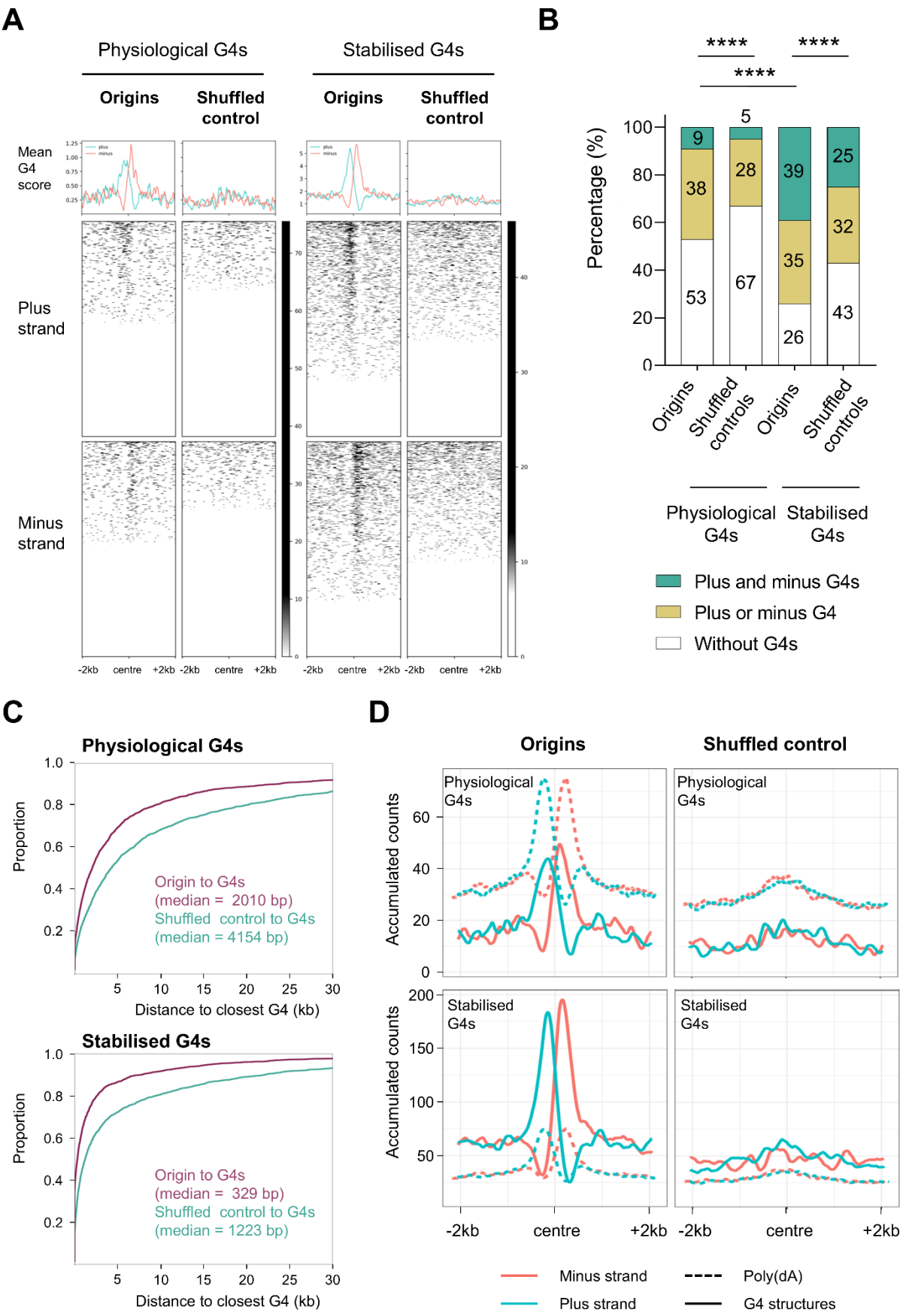


Figure 5. The distribution of G4 structures in the vicinity of origins.

A) The profile plots and heat maps show the distribution of the experimentally obtained G4 structures around centred origins and intergenic shuffled controls (±2 kb) in the Tb927 genomic background. The plus strand (light blue) and minus strand (pink) G4s, obtained under physiological conditions and in the PDS drug stabilized condition (72) were overlapped with the stranded SNS-seq mapped origins. Mean G4 score presents average G4 score (72) per 20 bp window. B) The proportions of origins that lack or possess G4 structures on one or both sides of the centre were determined. The ±2 kb window from the centre was analysed. Two sets of experimental G4 structures (72) were subjected to analysis in comparison with an intergenic shuffled control. The P-values were calculated using the Chi-square test, which involved a comparison of the absolute numbers of indicated categories for each pair of datasets (**** - P<0.0001). C) Empirical Cumulative Distribution Function (ECDF) of the distances between the origins and the closest physiological and stabilized G4 structures (72). Median distances are indicated. D) The profile plots illustrate the distribution of the poly(dA) sequence (AAAA) (dashed lines) and the experimental G4s (solid lines) around centred origins and intergenic shuffled controls (±2 kb). The analysis was conducted in the Tb927 genetic background. The plus strand is represented in blue and the minus strand in pink. The G4s were identified in the physiological and PDS drug stabilized condition (72). A smoothing function was employed to calculate the accumulated counts of the AAAA and G4s per position within a 100 bp window (50 bp up- and downstream from the position).

The origins exhibited varying nucleosome occupancy.

Previous studies demonstrated the presence of origins in nucleosome-depleted regions (34-37) and their positioning adjacent to well-positioned nucleosomes on one or both sides (18,35-38). Therefore, our subsequent objective was to determine whether there was a correlation between the mapped origins identified by stranded SNS-seq and the previously published MNase-seq datasets, which described the distribution of nucleosomes along the *T. brucei* genomic sequence (70). To achieve this, the origin and nucleosome positioning datasets were intersected, and a specific signature of nucleosomes was identified for the origins that was absent in the shuffled control. The origins exhibited a distinct "M" pattern of nucleosome distribution near the centre, with varying nucleosome occupancy (Figure 6A, Supplementary Figure 7A). The centre of origin displayed a markedly lower nucleosome occupancy (LNO) in comparison to the adjacent regions. Additionally, the centre of origin was situated between two regions exhibiting a higher nucleosome occupancy (HNO) in comparison to the shuffled control (Figure 6A, Supplementary Figure 7A). In light of the aforementioned evidence, our findings were found to be consistent with those previously documented in this field of study (18,34-38).

Subsequently, the number of origins without or with HNO region on one or both sides of the centre was quantified and compared to shuffled control (Materials and Methods). The results demonstrated that 64-67% of origins possessed an HNO region on one or both sides of the centre, compared to 51-54% of the shuffled controls (Figure 6B, Supplementary Figure 7B). The statistical analysis revealed a robust correlation between the origins and the presence of the HNO regions near the centres of origins (*P*-value = 6.13E-64) (Figure 6B).

Further investigation was conducted with the objective of determining the spatial relationship between poly(dA)/G4-enriched sequences and the nucleosome distribution. To achieve this, strand separated poly(dA) enriched sequences and the predicted G4s were overlapped with the MNase-seq data and centred around origins and intergenic shuffled controls. The plot

obtained following the intersections revealed that the LNO region was positioned between the plus and minus poly(dA)/G4 summits (Figure 6C). Additionally, the HNO peaks were found to exhibit overlap with the poly(dA) and G4 peaks and to surround the centre of origins (Figure 6C). Furthermore, these intersections permitted the estimation of the averaged distance of the above-mentioned elements from the centre of origins. The LNO region extended to approximately ±100 to 150 nt from the centre, thereby defining an area of approximately 200 to 300 bp in the centre of origin. The two HNO regions were located between 150 and 500 nt upstream and downstream from the centre of origin (Figure 6C).

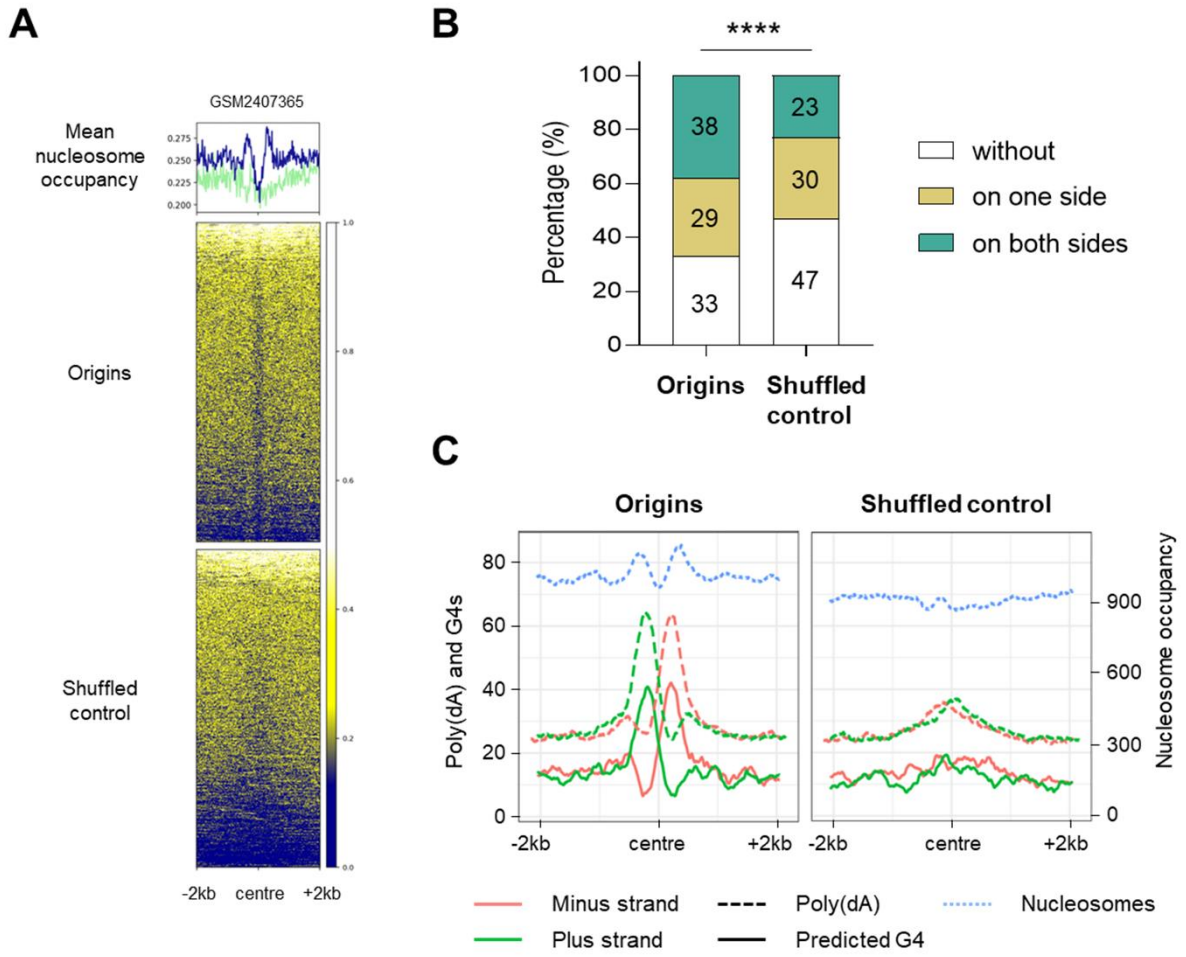


Figure 6. The distribution of nucleosomes around the origins in *T. brucei*.
A) The profile plot and heatmaps show the distribution of nucleosomes (70) around the centred origins (blue line) and the intergenic shuffled controls (green line) (± 2 kb). One replicate (GSM2407365) of MNase-seq data (70) was shown. Mean nucleosome occupancy presents average nucleosome score (70) per 20 bp window. B) The percentage of origins without or with high nucleosome occupancy (HNO) region on one or both sides of the centre compared to intergenic shuffled control. The *P*-values were calculated using the Chi-square test, which involved a comparison of the absolute numbers of indicated categories for each pair of datasets (**** - $P<0.0001$). C) The profile plots illustrate the distribution of the poly(dA) sequence (AAAA) (dashed line), the predicted G4 structures (solid line) and the nucleosomes distribution (70) (blue dotted line) around centred origins and intergenic shuffled controls (± 2 kb). The G4s were predicted by the G4Hunter application (73) in the Tb427 genomic background. The plus strand

is represented in green, and the minus strand in pink. The values on the plots present the accumulated counts of the AAAA, G4s and nucleosome occupancy scores per position within a 100 bp window (50 bp up- and downstream from the position).

The origins colocalize with intergenic R-loops enrichment sites.

R-loops are defined as three-stranded structures consisting of RNA-DNA hybrids and a displaced single-stranded DNA. It has been demonstrated that RNA-DNA hybrids occur naturally during cellular functions such as transcription and replication (79). However, the function of R-loops in the initiation of replication remained unclear (80,81). DRIP-seq is a widely used technique for genome-wide mapping of R-loops. This method is based on the immuno-precipitation of fragmented genomic DNA by the S9.6 monoclonal antibody, which specifically recognises RNA-DNA hybrids (82,83). A recent study employed DRIP-seq to map R-loops in *T. brucei*, revealing their enrichment within intergenic regions, between coding sequences of polycistronic transcription units, and coinciding with sites of polyadenylation and nucleosome depletion (71). Furthermore, DRIP-seq data from other model systems has indicated that long homopolymer poly(dA) tracks represent a significant contributor to R-loop formation (83,84). Considering the established role of short RNA primers in initiating DNA synthesis at the leading and lagging strands, an examination of the colocalization of the origins and previously published DRIP-seq data (71) was conducted. The result of the intersection showed an important accumulation of R-loops in the vicinity of origins, with the two upper summits situated near the origin centre (Figure 7A). An R-loop enrichment was also evident in the vicinity of the centre of the intergenic shuffled control (Figure 7A), which is consistent with previous reports indicating a high prevalence of R-loops in intergenic regions (71). However, this peak was observed to be lower in comparison to the peak observed near centres of origins. Furthermore, two upper summits were absent in the shuffled control. Subsequently, the number of origins and the number of shuffled controls that overlapped with R-loops were quantified. The results demonstrated that 90% of the origins and 68% of the shuffled controls exhibited overlap with R-loops (Figure 7B). Consequently, a statistically significant correlation was observed between the origins and R-loops (odds ratio = 4.17; *P*-value 2.87E-142), thereby establishing a clear connection between R-loop formation and origin activity.

Subsequently, the predicted G4 structures and poly(dA) enriched sequences from the plus and minus strands were superimposed with the DRIP-seq dataset (71) (Figure 7C) at the centres of origins and shuffled controls (Figure 7C). This intersection allowed us to estimate the relative positions of these elements and their averaged distance from the centre of origins. The two upper summits of R-loops were observed to be situated, on average, between the centre and 200 nt upstream and downstream from the centre of origins.

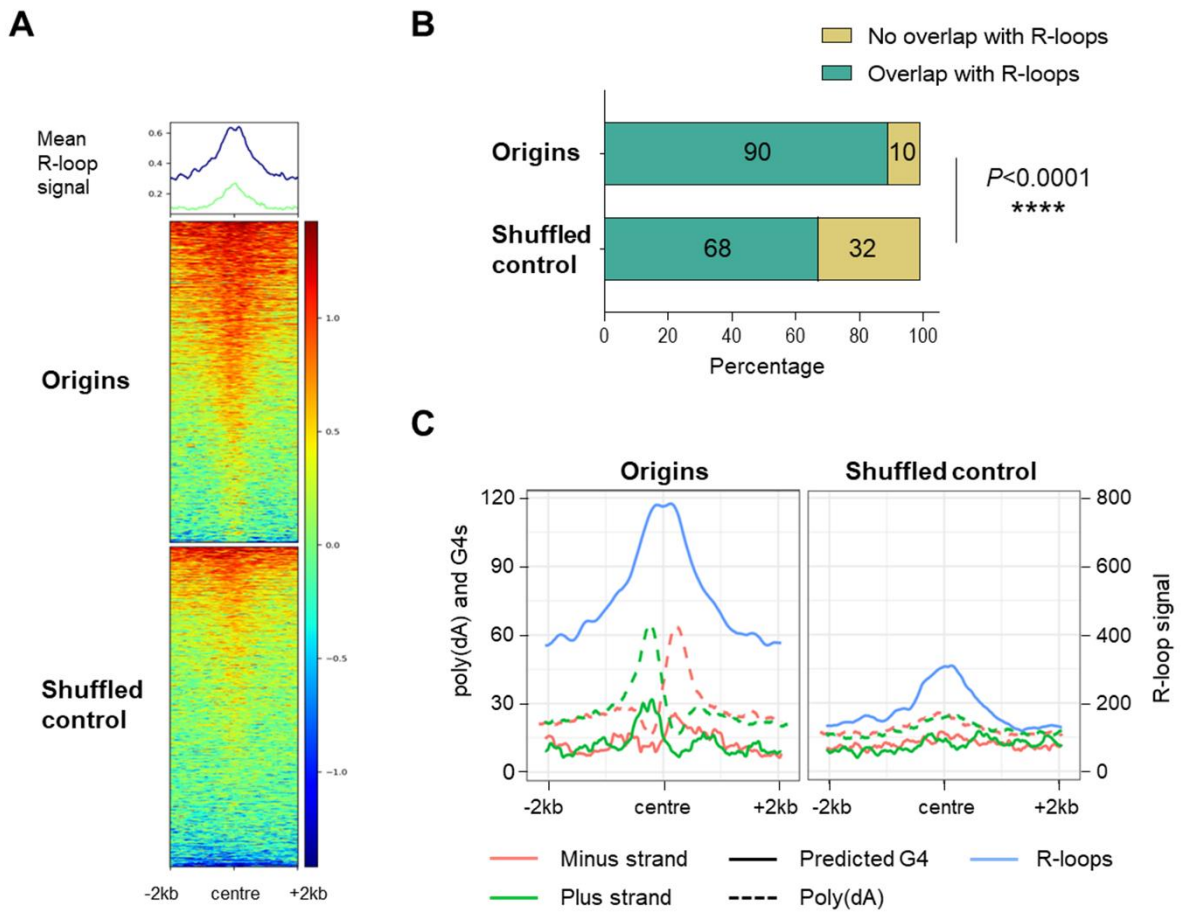


Figure 7. The distribution of R-loops around the origins.
A) The profile plots and heatmaps show the distribution of R-loops (71) around the centred origins (blue line) and the intergenic shuffled controls (green line) (+ 2 kb) in two cell types. Mean DRIP-seq signal presents average DRIP-seq signal (71) per 20 bp window. B) The proportion of origins and shuffled controls that overlap with R-loops. The *P*-values were calculated using Fisher's exact test, which involved a comparison of the numbers of indicated categories for each pair of datasets (**** - *P*<0.0001). C) The profile plots illustrate the distribution of the stranded poly(dA) sequence (AAAA) (dashed line), predicted G4s (solid line) and R-loops (71) (solid blue line) around centred origins and shuffled controls (± 2 kb). The plus strand is represented in green, and the minus strand in pink. The 13,409 G4s were predicted by the G4Hunter application (73) in the 427-2018 genomic background. The values on the plots present the accumulated counts of the AAAA, G4s and R-loop signals per position within a 100 bp window (50 bp up- and downstream from the position).

Modelling the topology of origins.
Only a limited number of studies have employed the stranded sequencing of short nascent strands for origin mapping (23,43). However, these studies have not provided an overview of the origin topography. The data obtained using the improved stranded SNS-seq protocol has prompted us to present a comprehensive model of a replication origin in *T. brucei*, with the averaged positions of a few genetic and epigenetic elements. According to our model, DNA replication starts at genomic sites where adenine and thymine are overrepresented and arranged in a specific pattern. On the plus strand, poly(dA) enriched sequences were situated

upstream, while poly(dT) enriched sequences were located downstream of the origin centre (Figure 8). Moreover, it was demonstrated that 47-74% of origins exhibited the presence of the experimental G4 structures within the ±2 kb region surrounding the centres (Figure 5B). Notably, the G4 structures near the centre of origins were embedded in poly(dA) rich sequences and the summits of G4 peaks covered 100-200 nt up and downstream from the centres (Figure 5D, Figure 8). Furthermore, our model incorporated data regarding the positions of nucleosomes and R-loops in the vicinity of origins. The centres of origin were identified as low nucleosome occupancy (LNO) regions with an average length of 200-300 nt. It was observed that these LNO regions were occasionally flanked by two high nucleosome occupancy (HNO) regions, which is likely due to the nucleosome sliding away from the centres of origin. The HNO regions were observed to occupy the regions from ±150 to ±500 bp on average, both upstream and downstream of the centre (Figure 8). R-loop enrichment was characterised by two distinct summits at the top of the R-loop peak, one upstream and one downstream of the centre, spanning a range from the centre to ±200 nt (Figure 8). Our model suggests that the leading strands of DNA replication, or SNS fibres, are synthesised using mainly the poly(dA)/G4-rich strands as templates.

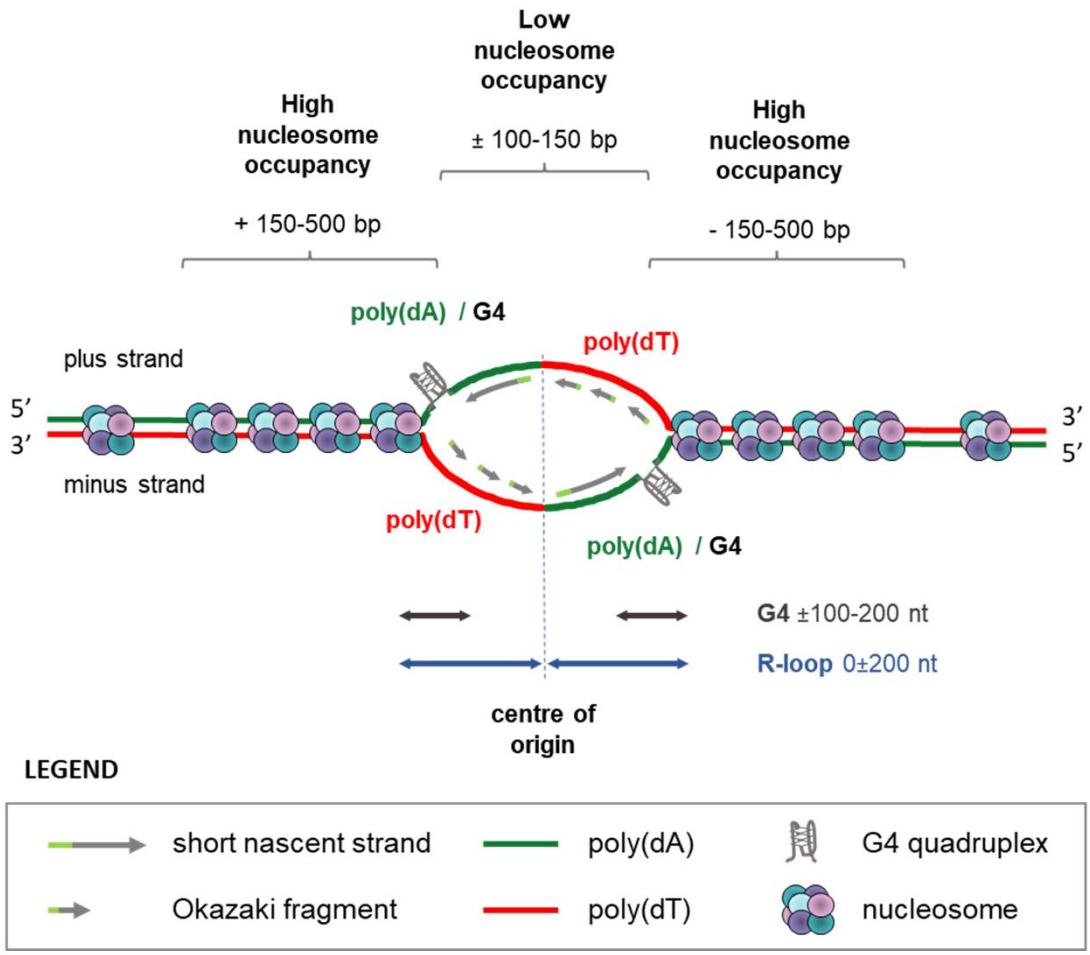


Figure 8. A proposed model of origin with the position of different genetic elements and nucleosome occupancy.

Poly(dA) enriched sequences, interspersed with G4 (poly(dA)/G4) are enriched upstream of the origin centres on plus strand and downstream of origin centres on minus strand. Poly(dT) enriched sequences are enriched downstream of origin centres on plus strand and upstream on minus strand. The centre of the origin is a low nucleosome occupancy (LNO) region, flanked by high nucleosome occupancy (HNO) regions. The double arrow lines indicate the position of the summits of the peaks of different origin elements. The presented positions were calculated from the averaged distances. It should be noted that not all origins have the same spacing. The G4 structures and HNO regions were identified at a limited number of origins; however, they are illustrated on this model to demonstrate the potential of origins to form these structures.

DISCUSSION

The genome of *T. brucei* is approximately one hundred times smaller than that of humans, which facilitates the study and mapping of DNA replication origins. To identify the active origins of this unicellular eukaryote we implemented stranded SNS-seq, which involves the isolation of RNA-primed short nascent leading strands (SNS) with preservation of SNS fibres orientation after sequencing. Our methodology is comparable to previously published origin-derived single-stranded DNA sequencing (Ori-SSDS) (23), but includes several differences. The stranded library preparation process was different in these two approaches and an SNS-depleted control, representing fibres resistant to λ-exonuclease treatments, was included in our study as a background control. The key benefit of stranded over conventional SNS-seq is the ability to identify origins with greater specificity by eliminating peaks that do not align with the proper orientation of SNS fibres of an active origin. Additionally, this method offers enhanced resolution and the potential to map the centres of origin or possible replication start sites.

In the present study, we mapped a set of active origins in *T. brucei*. However, a consensus sequence of mapped origins was not identified, which is consistent with the majority of studies conducted on other eukaryotes. Nevertheless, the genomic sites that function as origins are not random sequences. We have constructed a model of replication origin delineating the positions and orientations of several genetic elements and nucleosome positioning near origins (Figure 8). Our model indicates that the origins are situated between the poly(dA) and poly(dT) enriched sequences, exhibiting a distinctive nucleotide skew around the centre. The stranded SNS-seq approach has enabled us to successfully address the long-standing question in the field of DNA replication concerning the position and strand specificity of G4 structures near origins (27,32,37). Our research has demonstrated that the G4 structures were enriched near origins and intermingled with poly(dA) sequences. The G4 structures were formed on the strands that served as templates for the synthesis of the leading strands of replication. This finding counters the criticism that G4 enrichment around origins is an artefact resulting from inefficient λ-exonuclease digestion of the G4 secondary structure present at the 5′ end of SNS (42). Our data indicated that, in contrast to previous assumptions (42), SNS fibres do not contain G4 structures at their 5' end. Instead, they exhibited an enrichment in T nucleotides and were interspersed with C nucleotides at the 5' end.

The poly(dA)/G4 and poly(dT) rich DNA strands that constitute the origins have the potential to form complex secondary DNA structures that are expected to be dynamic and influenced by internal and external stimuli. However, by unifying these genetic elements, our study reconciles discrepant observations from literature concerning the structure of origins, which were previously characterized as either AT-rich (11-19) or GC-rich genomic regions, with the potential to form G4 structures (20-32). Our analysis indicated that G-rich sequences near origins tend to form G4 structures (Figure 5A, B, C, Supplementary Figure 6B). The detection of more G4s under PDS-stabilized conditions compared to physiological ones suggests that there is an important potential for G4s formation near origins. G4 formation seems to be very transient and therefore not easy to detect unless stabilized by drug. It seems that G4 structures near origins undergo rapid changes, either forming or dissolving in response to the replication dynamics. We therefore proposed that the formation of G4s plays a regulatory role in the activation of replication origins. Some of the recent publications suggested that G4s may serve as replication fork barriers and play a role in transient replication fork stalling (27,85). The enrichment of G4 structures on strands acting as templates for leading strand synthesis implies that the formation of G4 structures could potentially pose or arrest replication fork progression. A recent publication using a genetic approach has revealed that same-strand dimeric G4s were necessary and sufficient for replication initiation (37). Moreover, origin activity was found to be correlated with the formation of nucleosome free regions next to well-positioned nucleosomes (37). Given the estimated distances of the G4 summits, we suggested that the G4s could also serve as boundaries between low and high nucleosome occupancy regions.

Our study revealed that the centres of origin are significantly enriched with R-loops, and we disclosed for the first time a strong link between R-loop formation and DNA replication initiation. We think that a significant degree of overlap between data obtained by two distinct techniques – DRIP-seq and stranded SNS-seq – provides compelling validation of stranded SNS-seq origin mapping technique. We suggest that the observed enrichment of R-loops near the centre of origins may represent the DNA replication initiation sites. The R-loops are formed at initiation sites by RNA priming of SNS fibres and Okazaki fragments. However, Okazaki fragments are not exclusive markers of origins as they form R-loops throughout the genome, whereas R-loops formed by RNA-primed SNS are specific markers of origins. It can thus be proposed that the observed summits of R-loops situated adjacent to the centres of origins may be predominantly attributable to the enrichment of RNA-primed SNS fibres.

The results of our study demonstrated that most origins exhibited a poly(dA) and poly(dT) nucleotide skew near the centre. In addition, most origins overlap with R-loops situated near the centre. However, a subset of origins exhibited the presence of G4 structures near centre (9-39% origins possessed G4s on both sides, while 47-74% origins possessed G4s just on one side). Furthermore, a limited number of origins exhibited the presence of HNO regions near the centre (34-38% origins possessed HNO regions on both sides, while 64-67% origins possessed HNO region just on one side). Consequently, a subset of the mapped origins lacked any discernible G4 structures or HNO regions. This discrepancy from the proposed model may

be attributed to the fact that the intersected datasets were not obtained on the same cell strain. Furthermore, the cells were also grown and replicated under disparate conditions. An alternative hypothesis is that the efficiency of origins is regulated by the presence or absence of these elements. This would suggest that the most efficient origins exhibit the structure described in the model, while less efficient origins lack some of these elements. It can therefore be postulated that the presence or absence of these elements, in conjunction with the chromatin status and nuclear organisation of origins, serves to define the efficiency of origin activation.

The calculation of population-based IODs, performed in this study (Supplementary Figure 3C), revealed very long IODs that cannot be fully replicated during the previously estimated duration of S phase of 1.5-2.5 hours (86), pointing to the existence of additional origins in this parasite. Some weak or low-efficiency origins might be overlooked due to low activation at the cell population level, which resulted in low read coverage and subsequent loss after peak calling. Therefore, the number of origins of this parasite is likely to be higher than what has been reported here.

The genome wide origins and ORC-binding sites in *Plasmodium falciparum*, were described recently (45,87). The ORC-binding sites demonstrated a preference for areas of higher G/C content, an association with G4 forming structures (45,87). Despite *T. brucei* and *P. falciparum* belonging to disparate clades of divergent eukaryotes, our findings indicated that G4 structures were also present near origins of *T. brucei,* but predominantly within intergenic regions. This discrepancy could be explained by different genomic features of these two parasites. The overall A/T content in *P. falciparum* is over 80% and can rise to 95% in introns and intergenic regions, while the A/T content of the *T. brucei* genome is 56%. One study found that origins were preferentially activated in areas of low transcriptional activity (87) while the other study found an association of origins with gene bodies of strongly transcribed genes (45). In *T. brucei*, transcription is polycistronic (88,89), meaning that a set of genes is transcribed into one long mRNA that encodes several proteins after trans-splicing and polyadenylation (90). Our findings indicate that origins are predominantly localized in the intergenic regions within polycistronic transcription units (PTUs), which inevitably leads to collisions between DNA replication and transcription. The recent report indicated that origins were preferentially activated in areas of low transcriptional activity in *P. falciparum*, suggesting that the human malaria parasite has evolved to minimise conflicts between transcription and origin firing (87). Investigating how *T. brucei* avoid these collisions would be an interesting area for future research.

It is noteworthy that the origin structure described in this study in the early diverging eukaryote *T. brucei* exhibited a high degree of similarity to the described origin landscape in higher eukaryotes. The origins in mouse cells have been characterised as fragile sites, defined by long poly (dA:dT) tracks that are nucleosome free and devoid of the single-strand DNA-protecting protein RPA (19). Furthermore, the organisation of nucleosomes around the origins of *T. brucei* is identical to that observed in yeast and higher eukaryotes (18,34-38,91,92). This striking similarity indicates that the structure of origins may be a conserved feature throughout all

eukaryotes. However, further studies are necessary to gain deeper insights into the evolution of DNA replication origins.

DATA AVAILABILITY

The SNS-seq data (FASTQ files) from this study have been submitted to the European Nucleotide Archive under the accession number PRJEB70708. The SNS-seq data analysis pipeline is available at https://github.com/bridlin/finding_ORIs.

AUTHOR CONTRIBUTIONS

SS and YS designed the study. SS, LC and GS performed the experiments. BB developed the pipeline for the analysis of the stranded SNS-seq data. BB and PM analysed SNS-seq data. SS analysed combing data. YS provided funding. SS and BB wrote the initial draft of the manuscript. All authors edited and approved the final manuscript.

AKNOWLEDGEMENT

Special thanks to Cedric Notredame for advising us to prepare SNS stranded libraries. We would like to thank Julie A.J. Clement and Stephanie Le Gras for their help with preliminary analysis of SNS-seq data. We acknowledge the Montpellier DNA Combing Facility for providing silanized coverslips. We are grateful to MGX-Montpellier GenomiX (Montpellier, France) facility for performing sequencing. We would also like to acknowledge the 'Eukaryotic Pathogen, Vector and Host Informatics Resource' (VEuPathDB) for the integrated database, covering numerous eukaryotic pathogens (93-95). We are grateful to the Roscoff Bioinformatics platform ABiMS (http://abims.sb-roscoff.fr), part of the Institut Français de Bioinformatique (ANR-11-INBS-0013) and BioGenouest network, for providing computing resources. Thanks to Frédéric Bringaud (University Victor Segalen Bordeaux 2) for providing the *T. brucei* 427 wild-type PCF strain and Lucy Glover (Institute Pasteur, Paris) for providing us *T. brucei* 427 wildtype BSF strain.

FUNDING

The work was supported by the Agence Nationale de la Recherche (ANR) within the frame of the "Investissements d'avenir" program [ANR-11-LABX-0024-01 "PARAFRAP" to YS], the Centre National de la Recherche Scientifique (CNRS), the French Ministry of Research and the Centre Hospitalier Universitaire of Montpellier. SS was recipient of a grant from the Fondation pour la Recherche Médicale (ING20160435527). MGX acknowledges financial support from France Génomique National infrastructure, funded as part of "Investissement d'Avenir" program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09). ABiMS is founded as a part of the Institut Français de Bioinformatique (ANR-11-INBS-0013).

REFERENCES

1. Lee, C.S.K., Weiβ, M. and Hamperl, S. (2023) Where and when to start: Regulating DNA replication origin activity in eukaryotic genomes. *Nucleus*, **14**, 2229642.
2. Ganier, O., Prorok, P., Akerman, I. and Méchali, M. (2019) Metazoan DNA replication origins. *Curr Opin Cell Biol*, **58**, 134-141.
3. Hulke, M.L., Massey, D.J. and Koren, A. (2020) Genomic methods for measuring DNA replication dynamics. *Chromosome Res*, **28**, 49-67.
4. Leonard, A.C. and Méchali, M. (2013) DNA replication origins. *Cold Spring Harb Perspect Biol*, **5**, a010116.
5. Stinchcomb, D.T., Struhl, K. and Davis, R.W. (1979) Isolation and characterisation of a yeast chromosomal replicator. *Nature*, **282**, 39-43.
6. Brewer, B.J. and Fangman, W.L. (1987) The localization of replication origins on ARS plasmids in S. cerevisiae. *Cell*, **51**, 463-471.

7. Costa, A. and Diffley, J.F.X. (2022) The Initiation of Eukaryotic DNA Replication. *Annu Rev Biochem*, **91**, 107-131.
8. Bell, S.P. and Stillman, B. (1992) ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. *Nature*, **357**, 128-134.
9. Newlon, C.S. and Theis, J.F. (1993) The structure and function of yeast ARS elements. *Curr Opin Genet Dev*, **3**, 752-758.
10. Rowley, A., Cocker, J.H., Harwood, J. and Diffley, J.F. (1995) Initiation complex assembly at budding yeast replication origins begins with the recognition of a bipartite sequence by limiting amounts of the initiator, ORC. *EMBO J*, **14**, 2631-2641.
11. Okuno, Y., Satoh, H., Sekiguchi, M. and Masukata, H. (1999) Clustered adenine/thymine stretches are essential for function of a fission yeast replication origin. *Mol Cell Biol*, **19**, 6699-6709.
12. Kong, D. and DePamphilis, M.L. (2001) Site-specific DNA binding of the Schizosaccharomyces pombe origin recognition complex is determined by the Orc4 subunit. *Mol Cell Biol*, **21**, 8095-8103.
13. MacAlpine, D.M., Rodríguez, H.K. and Bell, S.P. (2004) Coordination of replication and transcription along a Drosophila chromosome. *Genes Dev*, **18**, 3094-3105.
14. Balasov, M., Huijbregts, R.P. and Chesnokov, I. (2007) Role of the Orc6 protein in origin recognition complex-dependent DNA binding and replication in Drosophila melanogaster. *Mol Cell Biol*, **27**, 3143-3153.
15. Stanojcic, S., Lemaitre, J.M., Brodolin, K., Danis, E. and Mechali, M. (2008) In Xenopus egg extracts, DNA replication initiates preferentially at or near asymmetric AT sequences. *Mol Cell Biol*, **28**, 5265-5274.
16. Segurado, M., de Luis, A. and Antequera, F. (2003) Genome-wide distribution of DNA replication origins at A+T-rich islands in Schizosaccharomyces pombe. *EMBO Rep*, **4**, 1048-1053.
17. Mojardín, L., Vázquez, E. and Antequera, F. (2013) Specification of DNA replication origins and genomic base composition in fission yeasts. *J Mol Biol*, **425**, 4706-4713.
18. Eaton, M.L., Galani, K., Kang, S., Bell, S.P. and MacAlpine, D.M. (2010) Conserved nucleosome positioning defines replication origins. *Genes Dev*, **24**, 748-753.
19. Tubbs, A., Sridharan, S., van Wietmarschen, N., Maman, Y., Callen, E., Stanlie, A., Wu, W., Wu, X., Day, A., Wong, N. *et al.* (2018) Dual Roles of Poly(dA:dT) Tracts in Replication Initiation and Fork Collapse. *Cell*, **174**, 1127-1142.e1119.
20. Delgado, S., Gómez, M., Bird, A. and Antequera, F. (1998) Initiation of DNA replication at CpG islands in mammalian chromosomes. *EMBO J*, **17**, 2426-2435.
21. Cadoret, J.C., Meisch, F., Hassan-Zadeh, V., Luyten, I., Guillet, C., Duret, L., Quesneville, H. and Prioleau, M.N. (2008) Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc Natl Acad Sci U S A*, **105**, 15837-15842.
22. Cayrou, C., Coulombe, P., Vigneron, A., Stanojcic, S., Ganier, O., Peiffer, I., Rivals, E., Puy, A., Laurent-Chabalier, S., Desprat, R. *et al.* (2011) Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. *Genome Res*, **21**, 1438-1449.
23. Pratto, F., Brick, K., Cheng, G., Lam, K.G., Cloutier, J.M., Dahiya, D., Wellard, S.R., Jordan, P.W. and Camerini-Otero, R.D. (2021) Meiotic recombination mirrors patterns of germline replication in mice and humans. *Cell*, **184**, 4251-4267.e4220.
24. Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., Marin, J.M. and Lemaitre, J.M. (2012) Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat Struct Mol Biol*, **19**, 837-844.
25. Valton, A.L., Hassan-Zadeh, V., Lema, I., Boggetto, N., Alberti, P., Saintomé, C., Riou, J.F. and Prioleau, M.N. (2014) G4 motifs affect origin positioning and efficiency in two vertebrate replicators. *EMBO J*, **33**, 732-746.
26. Cayrou, C., Ballester, B., Peiffer, I., Fenouil, R., Coulombe, P., Andrau, J.C., van Helden, J. and Méchali, M. (2015) The chromatin environment shapes DNA replication origin organization and defines origin classes. *Genome Res*, **25**, 1873-1885.
27. Comoglio, F., Schlumpf, T., Schmid, V., Rohs, R., Beisel, C. and Paro, R. (2015) High-resolution profiling of Drosophila replication start sites reveals a DNA shape and chromatin signature of metazoan origins. *Cell Rep*, **11**, 821-834.
28. Lombraña, R., Álvarez, A., Fernández-Justel, J.M., Almeida, R., Poza-Carrión, C., Gomes, F., Calzada, A., Requena, J.M. and Gómez, M. (2016) Transcriptionally Driven DNA Replication Program of the Human Parasite Leishmania major. *Cell Rep*, **16**, 1774-1786.

29. Langley, A.R., Gräf, S., Smith, J.C. and Krude, T. (2016) Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). *Nucleic Acids Res*, **44**, 10230-10247.
30. Sequeira-Mendes, J., Vergara, Z., Peiró, R., Morata, J., Aragüez, I., Costas, C., Mendez-Giraldez, R., Casacuberta, J.M., Bastolla, U. and Gutierrez, C. (2019) Differences in firing efficiency, chromatin, and transcription underlie the developmental plasticity of the. *Genome Res*, **29**, 784-797.
31. Akerman, I., Kasaai, B., Bazarova, A., Sang, P.B., Peiffer, I., Artufel, M., Derelle, R., Smith, G., Rodriguez-Martinez, M., Romano, M. *et al.* (2020) A predictable conserved DNA base composition signature defines human core DNA replication origins. *Nat Commun*, **11**, 4826.
32. Prorok, P., Artufel, M., Aze, A., Coulombe, P., Peiffer, I., Lacroix, L., Guédin, A., Mergny, J.L., Damaschke, J., Schepers, A. *et al.* (2020) Author Correction: Involvement of G-quadruplex regions in mammalian replication origin activity. *Nat Commun*, **11**, 3058.
33. Prioleau, M.N. and MacAlpine, D.M. (2016) DNA replication origins-where do we begin? *Genes Dev*, **30**, 1683-1697.
34. Lipford, J.R. and Bell, S.P. (2001) Nucleosomes positioned by ORC facilitate the initiation of DNA replication. *Mol Cell*, **7**, 21-30.
35. Berbenetz, N.M., Nislow, C. and Brown, G.W. (2010) Diversity of eukaryotic DNA replication origins revealed by genome-wide analysis of chromatin structure. *PLoS Genet*, **6**, e1001092.
36. Foss, E.J., Lichauco, C., Gatbonton-Schwager, T., Gonske, S.J., Lofts, B., Lao, U. and Bedalov, A. (2024) Identification of 1600 replication origins in. *Elife*, **12**.
37. Poulet-Benedetti, J., Tonnerre-Doncarli, C., Valton, A.L., Laurent, M., Gérard, M., Barinova, N., Parisis, N., Massip, F., Picard, F. and Prioleau, M.N. (2023) Dimeric G-quadruplex motifs-induced NFRs determine strong replication origins in vertebrates. *Nat Commun*, **14**, 4843.
38. Jansen, A. and Verstrepen, K.J. (2011) Nucleosome positioning in Saccharomyces cerevisiae. *Microbiol Mol Biol Rev*, **75**, 301-320.
39. Nasheuer, H.P. and Meaney, A.M. (2024) Starting DNA Synthesis: Initiation Processes during the Replication of Chromosomal DNA in Humans. *Genes*, **15**, 360.
40. Gerbi, S.A. and Bielinsky, A.K. (1997) Replication initiation point mapping. *Methods*, **13**, 271-280.
41. Bielinsky, A.K. and Gerbi, S.A. (1998) Discrete start sites for DNA synthesis in the yeast ARS1 origin. *Science*, **279**, 95-98.
42. Foulk, M.S., Urban, J.M., Casella, C. and Gerbi, S.A. (2015) Characterizing and controlling intrinsic biases of lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-quadruplex motifs around a subset of human replication origins. *Genome Res*, **25**, 725-735.
43. Senkevich, T.G., Bruno, D., Martens, C., Porcella, S.F., Wolf, Y.I. and Moss, B. (2015) Mapping vaccinia virus DNA replication origins at nucleotide level by deep sequencing. *Proc Natl Acad Sci U S A*, **112**, 10908-10913.
44. Kunnev, D., Freeland, A., Qin, M., Wang, J. and Pruitt, S.C. (2015) Isolation and sequencing of active origins of DNA replication by nascent strand capture and release (NSCR). *J Biol Methods*, **2**.
45. Castellano, C.M., Lacroix, L., Mathis, E., Prorok, P., Hennion, M., Lopez-Rubio, J.J., Méchali, M. and Gomes, A.R. (2024) The genetic landscape of origins of replication in P. falciparum. *Nucleic Acids Res*, **52**, 660-676.
46. Fenn, K. and Matthews, K.R. (2007) The cell biology of Trypanosoma brucei differentiation. *Curr Opin Microbiol*, **10**, 539-546.
47. Morrison, L.J., Steketee, P.C., Tettey, M.D. and Matthews, K.R. (2023) Pathogenicity and virulence of African trypanosomes: From laboratory models to clinically relevant hosts. *Virulence*, **14**, 2150445.
48. Cavalier-Smith, T., Chao, E.E., Snell, E.A., Berney, C., Fiore-Donno, A.M. and Lewis, R. (2014) Multigene eukaryote phylogeny reveals the likely protozoan ancestors of opisthokonts (animals, fungi, choanozoans) and Amoebozoa. *Mol Phylogenet Evol*, **81**, 71-85.
49. He, D., Fiz-Palacios, O., Fu, C., Fehling, J., Tsai, C. and Baldauf, S.L. (2014) An alternative root for the eukaryote tree of life. *Curr Biol*, **24**, 465-470.
50. Roger, A.J. (1999) Reconstructing Early Events in Eukaryotic Evolution. *Am Nat*, **154**, S146-S163.
51. Simpson, A.G. and Roger, A.J. (2004) The real 'kingdoms' of eukaryotes. *Curr Biol*, **14**, R693-696.
52. Stanojcic, S., Sollelis, L., Kuk, N., Crobu, L., Balard, Y., Schwob, E., Bastien, P., Pages, M. and Sterkers, Y. (2016) Single-molecule analysis of DNA replication reveals novel features in the divergent eukaryotes *Leishmania* and *Trypanosoma brucei* versus mammalian cells. *Sci Rep*, **6**, 23142.

53. Tiengwe, C., Marcello, L., Farr, H., Dickens, N., Kelly, S., Swiderski, M., Vaughan, D., Gull, K., Barry, J.D., Bell, S.D. *et al.* (2012) Genome-wide analysis reveals extensive functional interaction between DNA replication initiation and transcription in the genome of Trypanosoma brucei. *Cell Rep*, **2**, 185-197.
54. Devlin, R., Marques, C.A., Paape, D., Prorocic, M., Zurita-Leal, A.C., Campbell, S.J., Lapsley, C., Dickens, N. and McCulloch, R. (2016) Mapping replication dynamics in Trypanosoma brucei reveals a link with telomere transcription and antigenic variation. *Elife*, **5**.
55. Kim, H.S. (2019) Genome-wide function of MCM-BP in Trypanosoma brucei DNA replication and transcription. *Nucleic Acids Res*, **47**, 634-647.
56. Wirtz, E., Leal, S., Ochatt, C. and Cross, G.A. (1999) A tightly regulated inducible expression system for conditional gene knock-outs and dominant-negative genetics in Trypanosoma brucei. *Mol Biochem Parasitol*, **99**, 89-101.
57. Hirumi, H. and Hirumi, K. (1989) Continuous cultivation of Trypanosoma brucei blood stream forms in a medium containing a low concentration of serum protein without feeder cell layers. *J Parasitol*, **75**, 985-989.
58. Cayrou, C., Grégoire, D., Coulombe, P., Danis, E. and Méchali, M. (2012) Genome-scale identification of active DNA replication origins. *Methods*, **57**, 158-164.
59. Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins , L., Krueger, C. and Winget, S. (2010) FastQC A Quality Control tool for High Throughput Sequence Data. *Babraham Bioinformatics*, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
60. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10-12.
61. Langmead, B., Wilks, C., Antonescu, V. and Charles, R. (2019) Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*, **35**, 421-432.
62. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M. *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**.
63. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.
64. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011) Integrative genomics viewer. *Nat Biotechnol*, **29**, 24-26.
65. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F. and Manke, T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*, **44**, W160-165.
66. Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag, New York, book*.
67. Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, **2**, 28-36.
68. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H. and Glass, C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, **38**, 576-589.
69. Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res*, **14**, 1188-1190.
70. Maree, J.P., Povelones, M.L., Clark, D.J., Rudenko, G. and Patterton, H.G. (2017) Well-positioned nucleosomes punctuate polycistronic pol II transcription units and flank silent *Epigenetics Chromatin*, **10**, 14.
71. Briggs, E., Hamilton, G., Crouch, K., Lapsley, C. and McCulloch, R. (2018) Genome-wide mapping reveals conserved and diverged R-loop activities in the unusual genetic landscape of the African trypanosome genome. *Nucleic Acids Res*, **46**, 11789-11805.
72. Marsico, G., Chambers, V.S., Sahakyan, A.B., McCauley, P., Boutell, J.M., Antonio, M.D. and Balasubramanian, S. (2019) Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic Acids Res*, **47**, 3862-3874.
73. Brázda, V., Kolomazník, J., Lýsek, J., Bartas, M., Fojta, M., Šťastný, J. and Mergny, J.L. (2019) G4Hunter web application: a web server for G-quadruplex prediction. *Bioinformatics*, **35**, 3493-3495.
74. Techer, H., Koundrioukoff, S., Azar, D., Wilhelm, T., Carignon, S., Brison, O., Debatisse, M. and Le Tallec, B. (2013) Replication dynamics: biases and robustness of DNA fiber analysis. *J Mol Biol*, **425**, 4845-4855.
75. Fragkos, M., Ganier, O., Coulombe, P. and Mechali, M. (2015) DNA replication origin activation in space and time. *Nat Rev Mol Cell Biol*, **16**, 360-374.

76. Müller, L.S.M., Cosentino, R.O., Förstner, K.U., Guizetti, J., Wedel, C., Kaplan, N., Janzen, C.J., Arampatzi, P., Vogel, J., Steinbiss, S. *et al.* (2018) Genome organization and DNA accessibility control antigenic variation in trypanosomes. *Nature*, **563**, 121-125.
77. Valton, A.L. and Prioleau, M.N. (2016) G-Quadruplexes in DNA Replication: A Problem or a Necessity? *Trends Genet*, **32**, 697-706.
78. Masai, H. and Tanaka, T. (2020) G-quadruplex DNA and RNA: Their roles in regulation of DNA replication and other biological functions. *Biochem Biophys Res Commun*, **531**, 25-38.
79. Sollier, J. and Cimprich, K.A. (2015) Breaking bad: R-loops and genome integrity. *Trends in cell biology*, **25**, 514-22.
80. Lombraña, R., Almeida, R., Álvarez, A. and Gómez, M. (2015) R-loops and initiation of DNA replication in human cells: a missing link? *Front Genet*, **6**, 158.
81. Stuckey, R., García-Rodríguez, N., Aguilera, A. and Wellinger, R.E. (2015) Role for RNA:DNA hybrids in origin-independent replication priming in a eukaryotic system. *Proc Natl Acad Sci U S A*, -84.
82. Boguslawski, S.J., Smith, D.E., Michalak, M.A., Mickelson, K.E., Yehle, C.O., Patterson, W.L. and Carrico, R.J. (1986) Characterization of monoclonal antibody to DNA.RNA and its application to immunodetection of hybrids. *J Immunol Methods*, **89**, 123-130.
83. Wahba, L., Costantino, L., Tan, F.J., Zimmer, A. and Koshland, D. (2016) S1-DRIP-seq identifies high expression and polyA tracts as major contributors to R-loop formation. *Genes Dev*, **30**, 1327-1338.
84. Stanek, T.J., Cao, W., Mehra, R.M. and Ellison, C.E. (2022) Sex-specific variation in R-loop formation in Drosophila melanogaster. *PLoS Genet*, **18**, e1010268.
85. Zerbib, A. and Simon, I. (2023) Characterization of Unidirectional Replication Forks in the Mouse Genome. *Int J Mol Sci*, **24**.
86. da Silva, M.S., Muñoz, P.A.M., Armelin, H.A. and Elias, M.C. (2017) Differences in the Detection of BrdU/EdU Incorporation Assays Alter the Calculation for G1, S, and G2 Phases of the Cell Cycle in Trypanosomatids. *J Eukaryot Microbiol*, **64**, 756-770.
87. Totañes, F.I.G., Gockel, J., Chapman, S.E., Bártfai, R., Boemo, M.A. and Merrick, C.J. (2023) A genome-wide map of DNA replication at single-molecule resolution in the malaria parasite Plasmodium falciparum. *Nucleic acids research*, **51**.
88. Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., Lennard, N.J., Caler, E., Hamlin, N.E., Haas, B. *et al.* (2005) The genome of the African trypanosome *Trypanosoma brucei*. *Science*, **309**, 416-422.
89. Johnson, P.J., Kooter, J.M. and Borst, P. (1987) Inactivation of transcription by UV irradiation of *T. brucei* provides evidence for a multicistronic transcription unit including a VSG gene. *Cell*, **51**, 273-281.
90. Matthews, K.R., Tschudi, C. and Ullu, E. (1994) A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. *Genes Dev*, **8**, 491-501.
91. Segal, E. and Widom, J. (2009) Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr Opin Struct Biol*, **19**, 65-71.
92. Xu, J., Yanagisawa, Y., Tsankov, A.M., Hart, C., Aoki, K., Kommajosyula, N., Steinmann, K.E., Bochicchio, J., Russ, C., Regev, A. *et al.* (2012) Genome-wide identification and characterization of replication origins by deep sequencing. *Genome Biol*, **13**, R27.
93. Shanmugasundram, A., Starns, D., Böhme, U., Amos, B., Wilkinson, P., Harb, O., Warrenfeltz, S., Kissinger, J., McDowell, M., Roos, D. *et al.* (2023) TriTrypDB: An integrated functional genomics resource for kinetoplastida. *PLoS Negl Trop Dis*, **17**, e0011058.
94. Amos, B., Aurrecoechea, C., Barba, M., Barreto, A., Basenko, E., Bażant, W., Belnap, R., Blevins, A., Böhme, U., Brestelli, J. *et al.* (2022) VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic acids research*, **50**, D898-D911.
95. Warrenfeltz, S., Basenko, E., Crouch, K., Harb, O., Kissinger, J., Roos, D., A, S. and Silva-Franco, F. (2018) EuPathDB: The Eukaryotic Pathogen Genomics Database Resource. *Methods Mol Biol*, **1757**, 69-113.