Reviewed Preprint
v1 • December 4, 2025
Not revised

For correspondence:
slavica.stanojcic@umontpellier.fr
yvon.sterkers@umontpellier.fr

# These authors contributed equally

Competing interests: No competing interests declared

Funding: See page 25

Reviewing editor: Christine Clayton, Centre for Molecular Biology of Heidelberg University (ZMBH), Germany

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

Slavica Stanojcic[1,#], Bridlin Barckmann[1,#], Pieter Monsieurs[2], Lucien Crobu[1], Simon George[3], Yvon Sterkers[1]

[1]University of Montpellier, CNRS, IRD, Academic Hospital (CHU) of Montpellier, MiVEGEC, Montpellier, France • [2]Belgian Nuclear Research Centre (SCK•CEN), Mol, Belgium • [3]MGX-Montpellier GenomiX, University of Montpellier, CNRS, INSERM, Montpellier, France

eLife Assessment

The authors use sequencing of nascent DNA (DNA linked to an RNA primer, "SNS-Seq") to localise DNA replication origins in Trypanosoma brucei, so this work will be of interest to those studying either Kinetoplastids or DNA replication. The paper presents the SNS-seq results for only part of the genome, and there are significant discrepancies between the SNS-Seq results and those from other, previously-published results obtained using other origin mapping methods. The reasons for the differences are unknown and from the data available, it is not possible to assess which origin-mapping method is most suitable for origin mapping in T. brucei. Thus at present, the evidence that origins are distributed as the authors claim - and not where previously mapped - is **inadequate**.

https://doi.org/10.7554/eLife.108143.1.sa2

Abstract

The universal features that define genomic regions acting as replication origins remain unclear. In this study, we mapped a set of origins in *Trypanosoma brucei* using stranded short nascent strand sequencing method. Our results showed that DNA replication predominantly initiates in intergenic regions between poly(dA)- and poly(dT)-enriched sequences. G4 structures were detected in the vicinity of some origins and were embedded in poly(dA)-enriched sequences in a strand-specific manner: G4s on the plus strand were located upstream, while those on the minus strand were located downstream of the centre. The origins' centres were found to be areas of low nucleosome occupancy, surrounded by regions of high nucleosome occupancy. Furthermore, our results demonstrate that 90% of replication origins overlap with a minor proportion of the previously reported R-loops. These findings shed new light on the sequence and structural features that define the topology of replication origins in *T. brucei*. To further characterize replication dynamics at the single-molecule level, we employed DNA combing analysis.

Introduction

DNA replication is one of the most fundamental processes by which a cell creates an exact copy of its genome prior to 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 [1–3]). The first isolated and characterized eukaryotic origin comes from budding yeast, *Saccharomyces cerevisiae*, where DNA replication starts from the AT-rich autonomously replicating sequence (ARS) with a loosely defined and thymidine-rich ARS consensus sequences (ACS) [4]. The other eukaryotic origins lack a unique

consensus sequence. Some studies reported increased AT-content [5–11] or long poly(dA:dT) tracts as replication origins [12, 13]. Other reports described the origins as regions of a high GC-content, such as CpG islands [14–17] or regions near G-rich sequences with the potential to form G-quadruplexes (G4s) [18–26]. In addition, several epigenetic factors and chromatin environments were linked to eukaryotic origins, but without a common signature [1,22,27]. However, replication origins exhibit several characteristics that distinguish them from the surrounding chromatin. They appear in nucleosome-depleted regions [28–31], which are flanked on one or both sides by well-positioned nucleosomes [12,29–32]. An active origin is characterized by the generation of two divergent single-stranded DNA molecules, each with a short RNA primer at 5' end, known as short nascent strands (SNS) (reviewed in [33]). Size-selected SNS can be enriched and purified from DNA molecules 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 [34]. Thereafter, this technique has been successfully applied in different species, using approaches such as SNS-on-chip [15, 16] or by high throughput sequencing of SNS (SNS-seq) [18, 20, 22, 25, 26, 35, 36]. A recent development has improved the conventional SNS-seq method by enabling the preservation of SNS molecule directionality following sequencing [17, 37].

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 (PCF) in the midgut of the tsetse fly [38]. The order *Trypanosomatida* diverged independently of the *Opisthokonta*, the group comprising animals and fungi [39], and has attracted attention as an original model for biological and comparative studies. These single-celled eukaryotes have some unusual genomic and epigenetic properties. Transcription is polycistronic, meaning that multiple functionally unrelated genes are transcribed into one long precursor mRNA [40]. A better understanding of the fundamental processes in these divergent eukaryotes could provide insight into evolutionary steps and pave the way for innovative medical treatments for sleeping sickness and nagana disease.

The replication of DNA in these parasites has recently become a subject of interest, [41, 42, 43]. To map and identify active DNA replication origins (hereafter referred to as "origins") in *T. brucei*, we isolated RNA-primed SNS molecules and sequenced them using a stranded protocol that preserves their orientation. This stranded SNS-seq method allowed the mapping of origins between two divergently oriented SNS peaks and the discovery of potential DNA replication start sites. We then conducted an extensive bioinformatics analysis of the mapped origins and compared our findings with previously published scientific datasets. This enabled us to identify several key components of origins and to propose a model of an origin in *T. brucei*.

Results

Origin mapping by stranded SNS-seq

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. SNS-enriched samples were prepared through a two-stage process. SNS molecules were enriched first, by size-selection (0.5-2.5 kb), and second, by removing broken DNA molecules by λ-exonuclease treatments. SNS-depleted controls, prepared in parallel with SNS-enriched samples, underwent an additional RNase treatment to remove RNA primers before λ-exonuclease treatment (Methods and Figure 1A). These contained molecules resistant to RNase and/or λ-exonuclease and served as background controls. Then, stranded libraries were prepared for high-throughput sequencing (Figure 1B and Methods). Paired-end sequencing of an SNS-enriched samples and an SNS-depleted controls was performed on three biological replicates of PCF and BSF cells. Read counts after sequencing and each analysis step are shown for each sample in Supplementary Table 1 and Figure S1A.

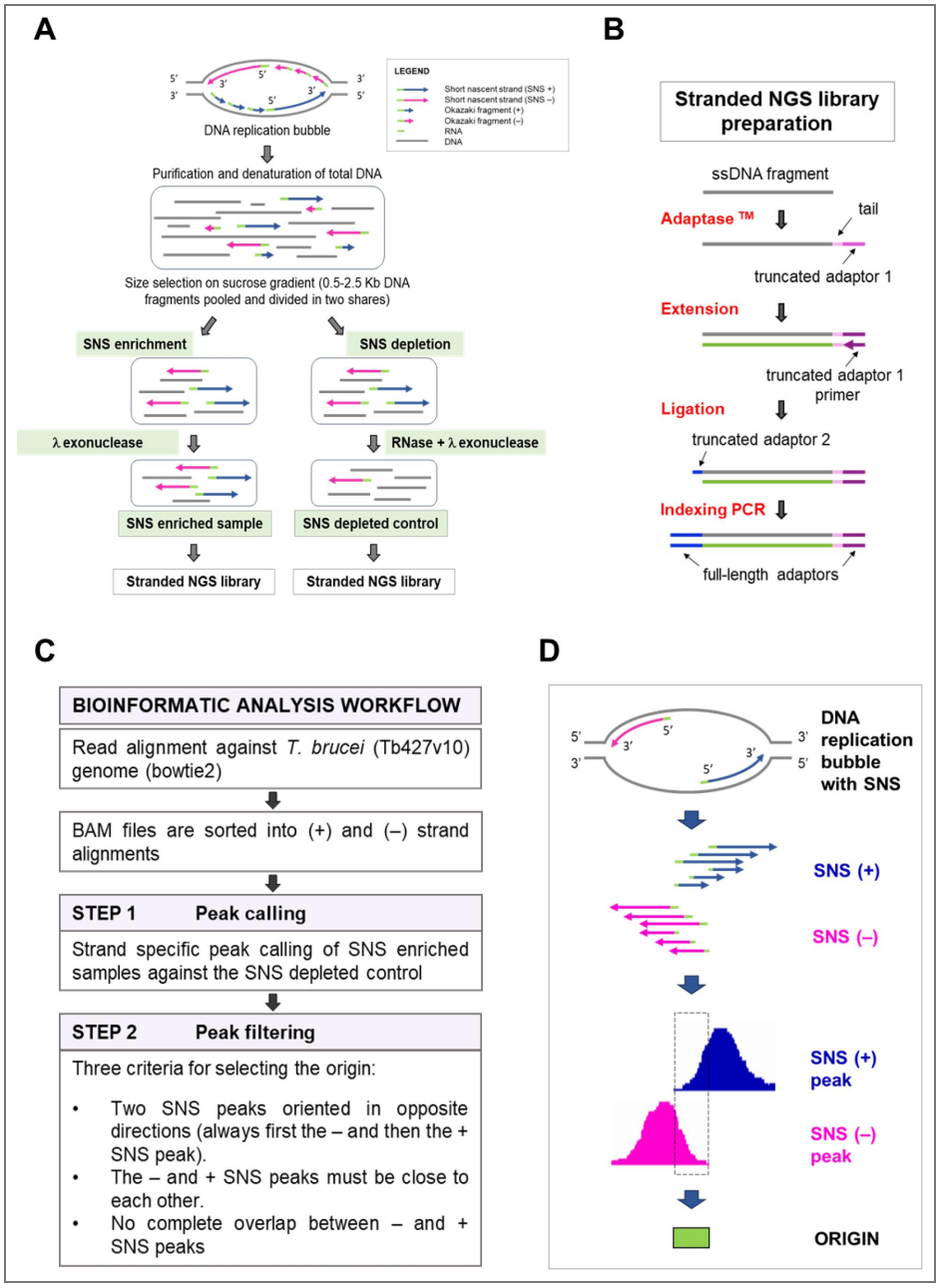

Figure 1. Experimental protocol and bioinformatics workflow for selecting the origins.

A) Schematic representation of the experimental workflow for short nascent strand (SNS) purification. A replication bubble with 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) Workflow of stranded library preparation for Illumina sequencing. Single stranded DNA fragments were tailed and tagged at the 3' end with truncated adaptor 1. Primer extension was used to generate the complementary strand. The ligation step was used to add the truncated adaptor 2. The indexing PCR step included full-length adaptors. C) 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. D) The schematic representation shows one origin, generating a replication bubble with two SNS leading strand in divergent orientation (SNS + and SNS –). The plus and minus SNS are presented to illustrate 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 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.

The bioinformatics workflow used for origin selection is schematically presented in Figure 1C and Methods. In brief, after the alignment and strand-specific sorting of the reads, two steps of peak selection were performed. The first step was strand-specific peak calling against SNS-depleted control. The second step, called "peak filtering", identified peaks pairs consistent with the divergent orientation of SNS molecules at active origins (Figure 1C). Therefore, only peak pairs with a minus peak in front of a plus peak, and both in proximity (Methods and Figure 1C and 1D) were retained as positives, while 88-96% of the called peaks were removed in the "peak filtering" step (Supplementary Figure 1B and 1C). Here, we used information about the divergent orientation of SNS, a key feature of active origins, to increase the specificity of origin mapping. Ultimately, the origins were identified as genomic regions situated between the inner borders of filtered peak pairs (Figure 1D).

A representative screenshot of the distribution of plus and minus peaks and the positions of the mapped origins for three replicates of BSF cells is shown in Figure 2A. We found a variable number of origins in the replicates (Supplementary Figure 1D), and the overlap of origins between replicates varied from 8-64%. The distribution of the lengths of the mapped origins was estimated for each replicate (Supplementary Figure 2A), and the average length of origins was found to be approximately 150 bp. 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. To identify preferred genomic sites of origin activation, we analysed origin distribution around centred intergenic and genic regions. Origins were unevenly distributed, showing higher density in intergenic and lower in genic regions (Figure 2B). This intergenic enrichment was consistent across both cell types and absent in shuffled controls, which consists of randomly selected regions matched for size, number, and chromosomal distribution (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 (Figure 2C). This distribution was significantly different from the distribution that would be expected if the origins were distributed evenly across the genome (*P*-value<0.0001) (Figure 2C).

We then wanted to see if origins were organised as solitary sites or grouped together, so we examined how they clustered within genomic regions of increasing length. Our results showed that *T. brucei* origins were predominantly organised in clusters, with 85.8% of PCF origins and 83.2% of BSF origins organized in clusters of 50 kbp (Supplementary Figure 2B). Empirical cumulative distribution function (ECDF) was calculated for both cell types, and a statistically significant difference was confirmed between the origins and shuffled controls clustering (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 means that origins tend to be activated within initiation zones or clusters rather than as solitary initiation sites in *T. brucei*. This finding is consistent with previous research in other organisms, which has shown that origins are unevenly distributed and tend to form clusters of different lengths [20, 22, 25]. Next, the intersection of the two sets of origins revealed that 45% of origins were common between the PCF and BSF cell types (Supplementary Figure 2C). The number of origins in both cells was compared using normalised mapped reads (Methods), and after merging the reads from the three replicates and normalization, we showed that BSF cells contained almost twice as many origins as PCF cells (merged BSF: 844 origins and averaged normalised subsamples of PCF: 454 origins) (Supplementary Figure 2D).

We then calculated the inter-origin distances (IODs) distribution at the population level 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 2E). Very long IODs were rarely observed, with maximum distances of 354 and 488 kilobases between PCF and BSF origins, respectively (Supplementary Figure 2E).

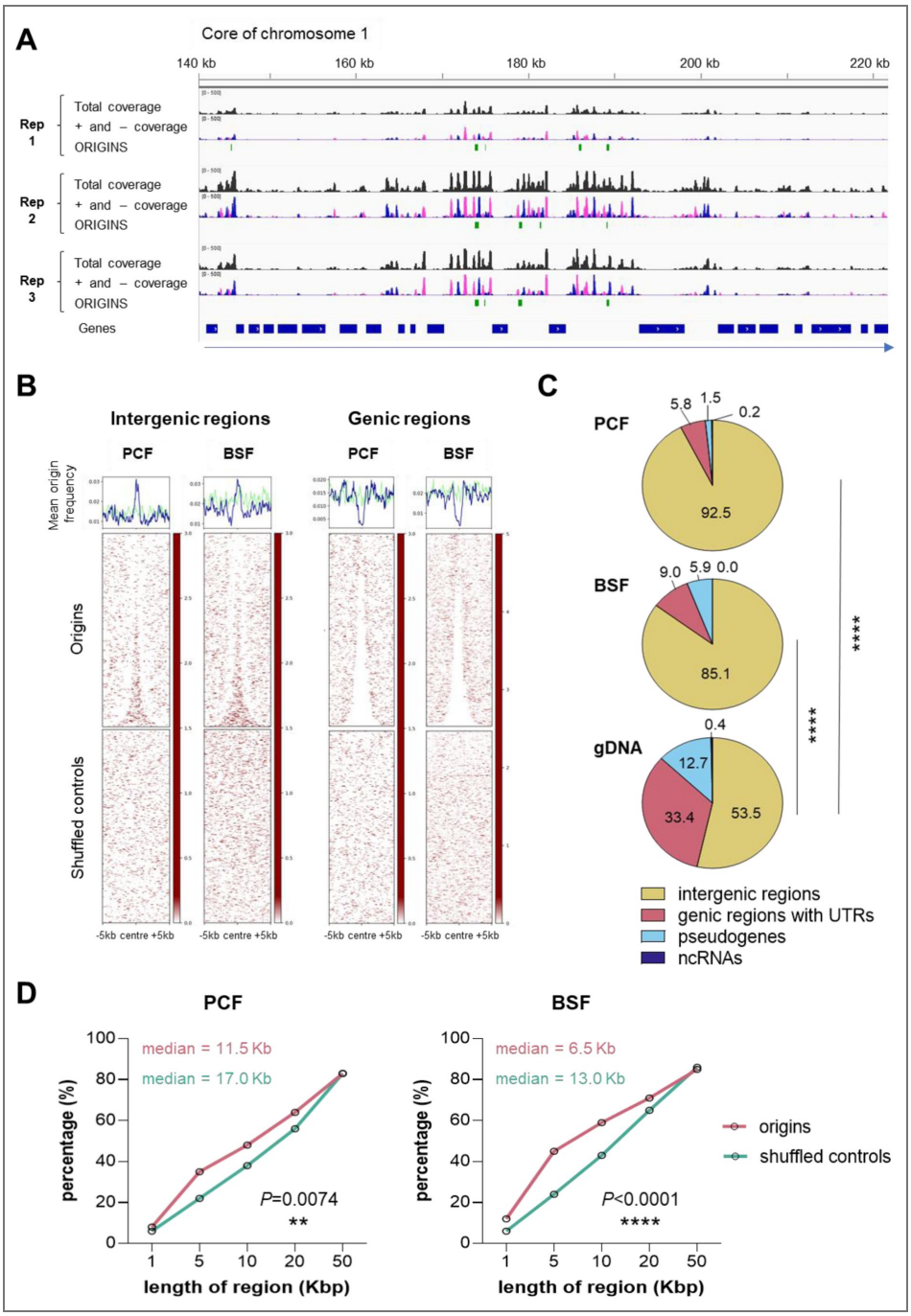

Figure 2. Genomic distribution of mapped origins.

A) A representative screenshot from Integrative Genomics Viewer (IGV) presenting 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. The blue bars show the position of the genes and the direction of transcription of the polycistronic unit is indicated by the blue arrow below. 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) Empirical cumulative distribution function (ECDF) showing the proportion of pairwise distances between origins (pink) or between shuffled control regions (green). The *P*-values were computed using Chi-square tests.

Insights from DNA molecular combing: comparative analysis of single-molecule DNA replication and cell population origin use

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 (reviewed in [44]). DNA combing provides an overview of DNA replication parameters at the single-molecule 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 (Methods). After immuno-detection, the three discrete signals were discerned on combed DNA. The red signal, followed by green, indicated the direction of replication forks, while the blue signal represented the unmodified DNA (Figure 3A). Origins were positioned at the centre of two divergently oriented replication forks. DNA replication parameters, including replication fork speed, IODs, and fork asymmetry, were determined as previously described [41]. Measurements from three biological replicates were pooled and column statistics are shown in Supplementary Table 2. The median replication fork velocity was 1.73 and 2.44 kb/min 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, with these differences between the two cell types being statistically significant (Figure 3B and Figure 3C). Moreover, we compared the frequency of asymmetric replication forks, defined as bidirectional forks that progress from the 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 [45], 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 DNA molecules 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 DNA molecules 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). The stranded SNS-seq method revealed 549 origins for PCF and 914 origins for BSF under stringent origin selection, which represent a fraction of all origins. Despite the inability of either method to determine the exact number of origins, the number of origins in a single cell tends to be smaller than in a population of cells. This phenomenon, known as flexible origin usage, is a well-documented feature of eukaryotic origins (reviewed in [44]).

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

Having mapped a subset of origins of PCF and BSF cells, our next objective was to identify the common elements that would define the origins of *T. brucei*. To do so, we merged the origins of PCF and BSF cells into a single set, bringing the total number of origins to 1225. All analyses were also performed on the separate PCF and BSF sets of origins to demonstrate the similarity of their structures, and the results are presented in the supplementary figures. As most origins were identified within intergenic regions, another shuffled control was developed to represent random intergenic regions selected in accordance with the number of origins, their size, and chromosomal location (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. Our first objective was to examine the nucleotide distribution of centred origins and compared it with the centred 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

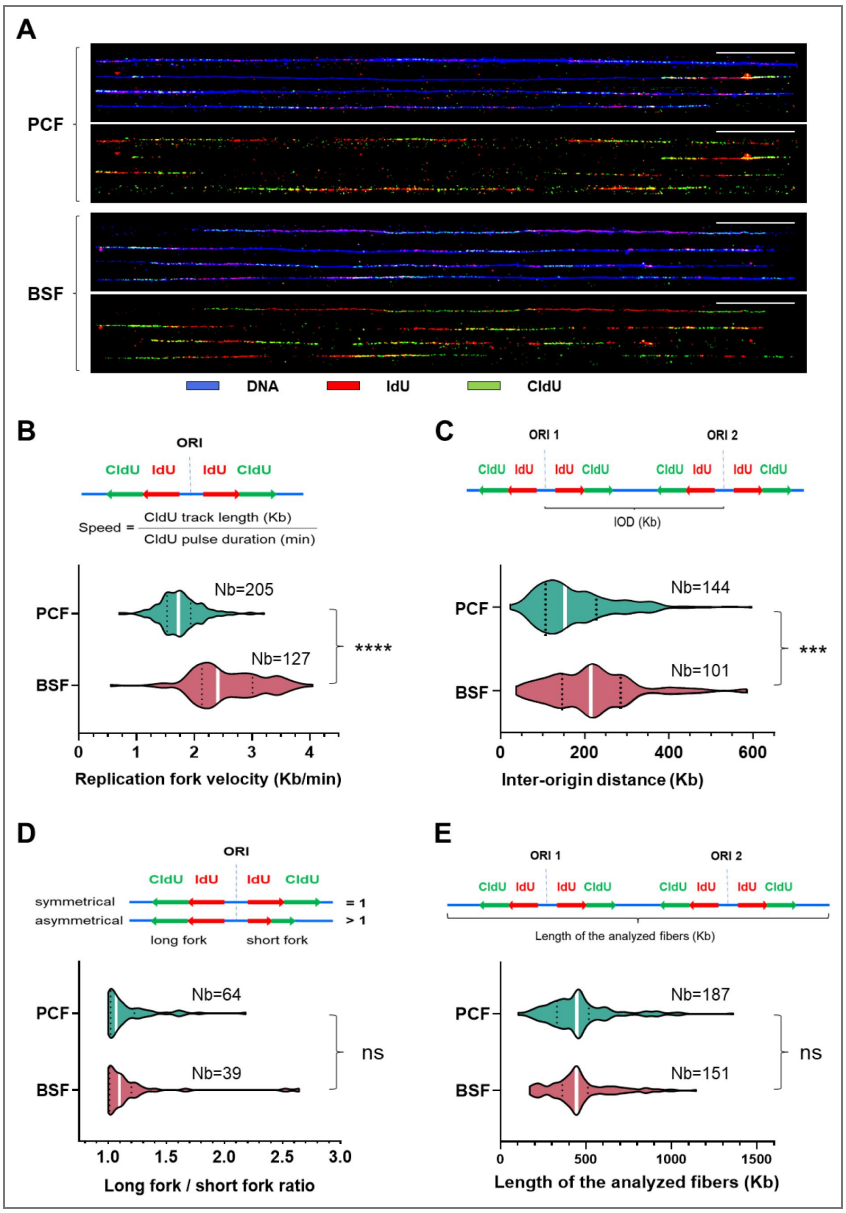

Figure 3. Comparison of DNA replication parameters between PCF and BSF cells by DNA molecular combing.

A) The figure depicts representative DNA molecules after immuno-detection. The immuno-detected DNA is 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 red and green tracks from the first and second pulse of nucleosides extracted from the corresponding upper panels. 50 kb scale bars are shown as white lines. 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 intact DNA molecules (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 [41]. 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 analysed DNA lengths. The lower panel presents violin plots with the measured lengths of the analysed DNA 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; * - $P<0.05$; ** - $P<0.01$; *** - $P<0.001$; **** - $P<0.0001$). The number of measurements (Nb) is given for each cell type.

intergenic regions (reaching 30%) (Figure 4A), while the overall A or T content of the genomic sequence is approximately 28%. 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 3A). 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 3A). The subsequent step was to search for consensus motives of origins (Methods). However, apart from extended sequences that were enriched with adenines or thymines no discernible consensus motif was identified (data not shown). The sequences surrounding the origins were also analysed using the WebLogo software [46], 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 origins (Supplementary Figure 3B). We then looked at the distribution of polynucleotides of different lengths around the origins. Poly(dA)-rich sequences were enriched upstream and poly(dT)-rich sequences downstream of the origins, a pattern absent in shuffled controls (Figure 4B, Supplementary Figure 3C). Interestingly, the upstream region showed an enrichment of up to four consecutive Gs, while the downstream region showed an enrichment of up to four consecutive Cs (Figure 4B). We also checked nucleotide and polynucleotide distributions around PCF and BSF origins separately and found that the distributions were similar in both cell types and equivalent to the merged set of origins (Supplementary Figure 4A and Supplementary Figure 4B). Accordingly, an origin can be defined as a genomic site with a specific uneven distribution of nucleotides, 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 controls (Figure 4C).

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

G4s, DNA secondary structures with non-canonical Watson-Crick base pairing, have recently emerged as potential regulators of genome replication [47]. Given the previously reported presence of G4 structures in the vicinity of some eukaryotic origins [18, 19, 21, 26], we investigated the correlation between G4s and origins of *T. brucei*. To do this, we used two sets of G4s that had previously been determined by G-quadruplex sequencing (G4-seq) [48]. 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 G4s [48]. The SNS-seq origin dataset was intersected with stranded G4-seq datasets obtained in both conditions (Methods), resulting in a notable enrichment of G4s in the vicinity of the origins (Figure 5A and Supplementary 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 and Supplementary Figure 5A). This pattern of G4 enrichment was absent in the intergenic shuffled controls (Figure 5A and Supplementary Figure 5A). The occurrence of

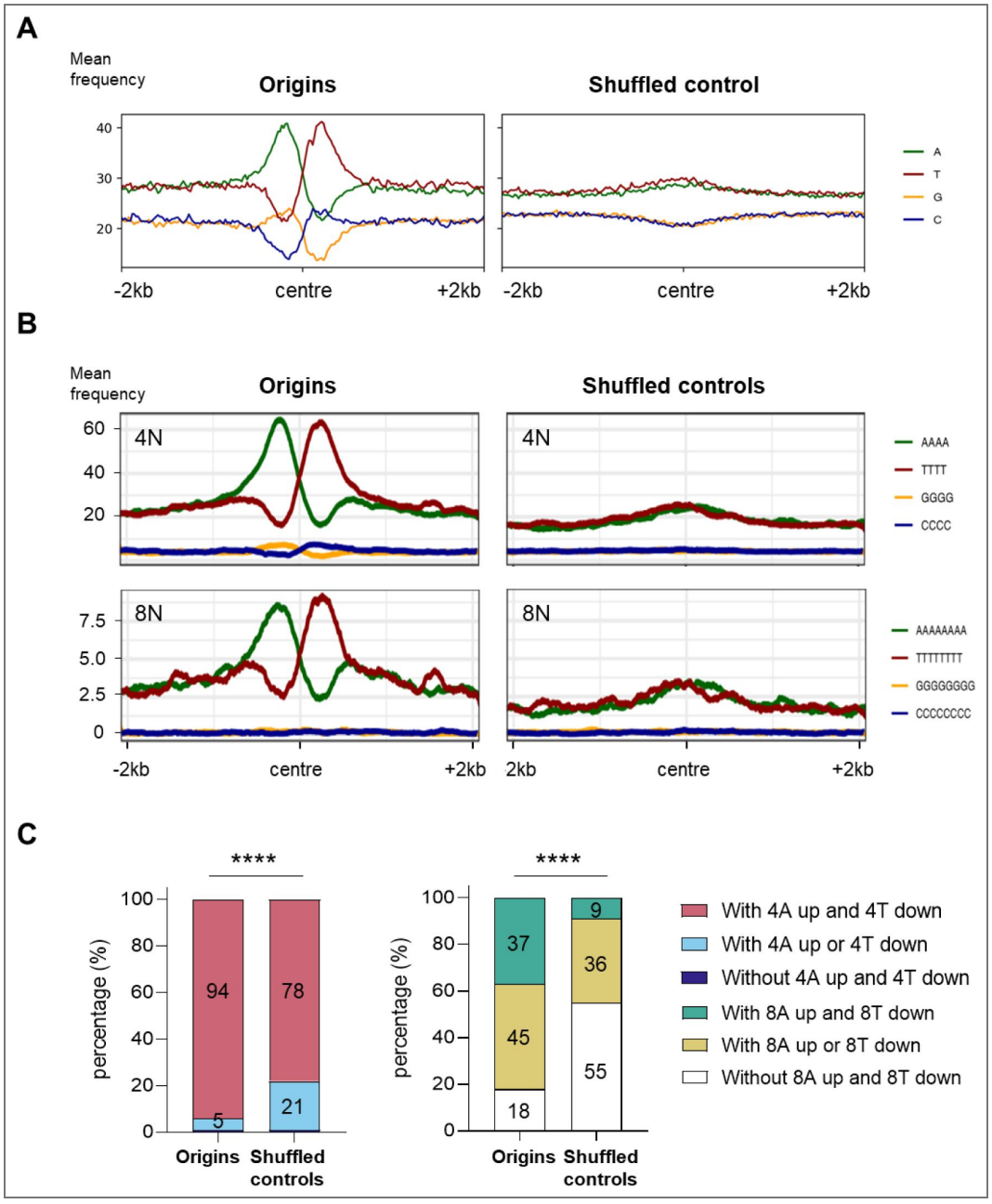

Figure 4. Spatial organisation of nucleotides and polynucleotides in the vicinity of origins.

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 (**** - $P<0.0001$).

G4s near the origins varied, with a discernible increase of G4s near the centre observed under PDS-stabilised conditions in comparison to physiological conditions (Figure 5A and Supplementary 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 5B). Quantification of G4 distribution around the PCF and BSF origins, was similar in the two cell types, and comparable to the merged set of origins (Supplementary Figure 5C).

To gain a deeper insight into the distribution of origins and the G4 structures, an empirical cumulative distribution function (ECDF) was 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). Half of the origins had a stabilised G4 structure within ±329 bp of the centre, while 80% of the origins had the same structure within ±2575 bp of the centre. Therefore, 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 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 strand separated experimental G4s [48] were intersected at centred origins and shuffled controls (Methods). 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 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.

The origins exhibited varying nucleosome occupancy

Previous studies demonstrated the presence of origins in nucleosome-depleted regions [28–31] and their positioning adjacent to well-positioned nucleosomes on one or both sides [12, 29–32]. Therefore, our subsequent objective was to determine whether there was a correlation between the origins identified by stranded SNS-seq and the previously published MNase-seq datasets, which described the distribution of nucleosomes in *T. brucei* [49]. After intersecting the origins and nucleosome positioning datasets (Methods), we identified a nucleosome signature specific to origins that was absent in the shuffled control. Origins showed a distinct “M” pattern of nucleosome distribution around the centre, with varying nucleosome occupancy (Figure 6A and Supplementary Figure 6A). 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 controls (Figure 6A and Supplementary Figure 6A).

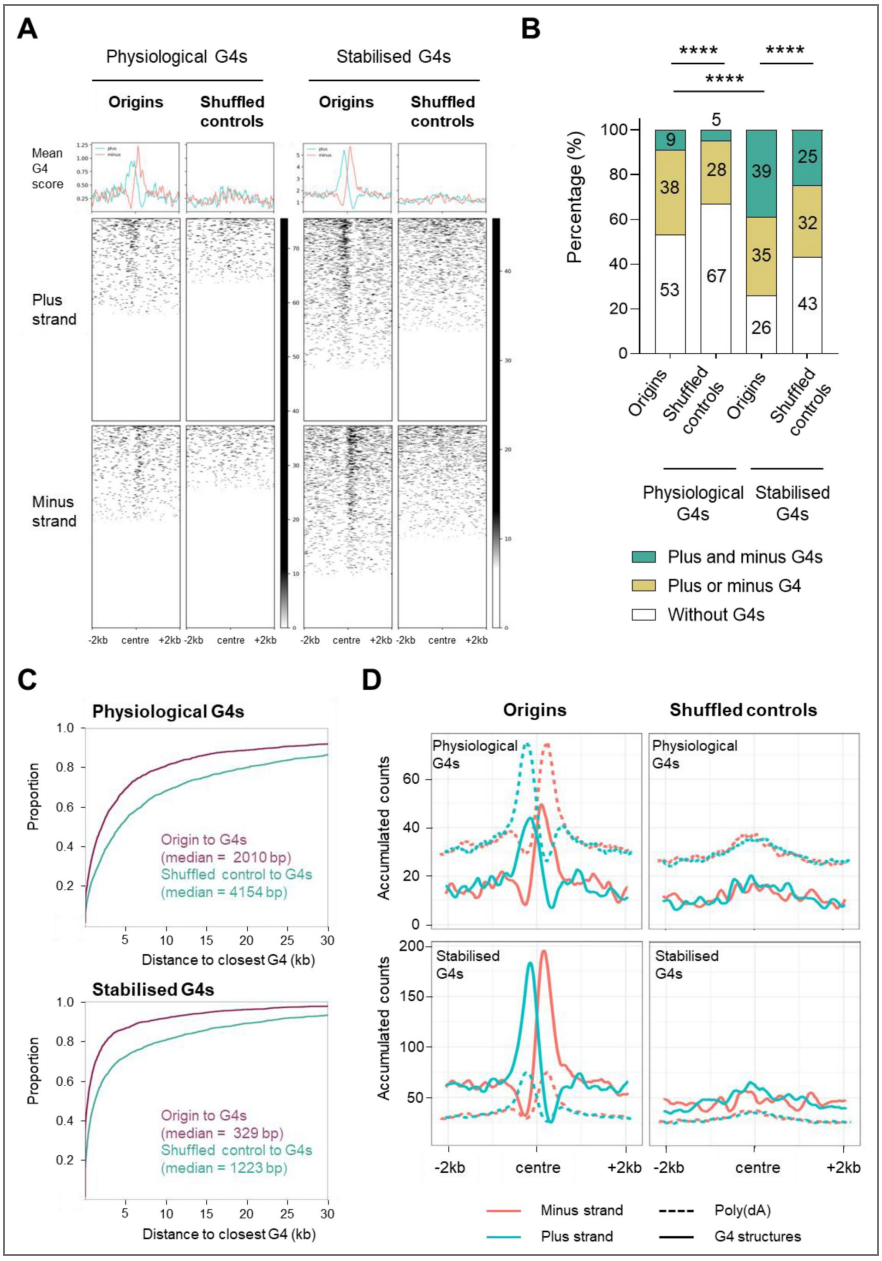

Figure 5. Distribution of G4 structures in the vicinity of origins.

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 Tbb TREU927 reference genome (Methods). The plus strand (light blue) and minus strand (pink) G4s, obtained under physiological conditions and in the PDS drug stabilized condition [48] were overlapped with origins mapped by stranded SNS-seq. Mean G4 score presents average G4 score [48] 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 for the presence of G4s. Two sets of experimental G4 structures [48] were subjected to analysis in comparison with intergenic shuffled controls. 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 [48]. Median distances are indicated. D) The profile plots illustrate the distribution of the poly(dA) sequence (AAAA) (dashed lines) and the experimental G4s (physiological and PDS drug stabilized) [48] (solid lines) around centred origins and intergenic shuffled controls (±2 kb). The analysis was conducted in the Tbb TREU927 reference genome (Methods). The plus strand is represented in blue and the minus strand in pink. 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).

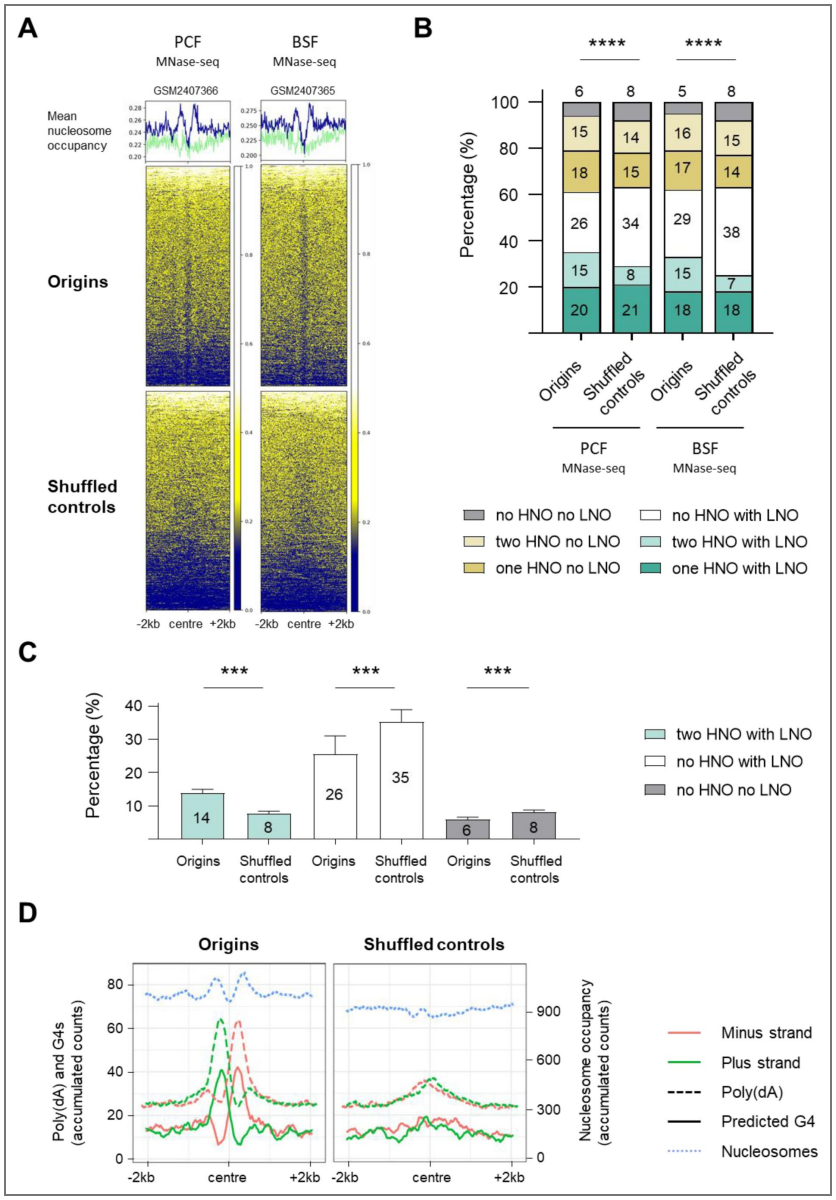

Figure 6. The distribution of nucleosomes around the origins in *T. brucei*.

A) The profile plot and heatmaps show the distribution of nucleosomes (MNase-seq data) detected in PCF and BSF cells [49] around the centred origins (blue line) and the intergenic shuffled controls (green line) (± 2 kb) in Tb Lister 427 reference genome. One replicate of MNase-seq data [49] for PCF (GSM2407366) and one replicate for BSF cells (GSM2407365) is shown here, the other replicates are presented in Supplementary Figure 6A. Mean nucleosome occupancy presents average nucleosome score (dyad value) [49] per 20 bp window. B) The percentage of origins and intergenic shuffled controls with or without high nucleosome occupancy (HNO) and low nucleosome occupancy (LNO) regions in the indicated combinations. Quantification was performed for the same replicates as indicated in the Figure 6A. 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) Percentage of the three specified categories of HNO and LNO combinations that were significantly different between origins and shuffled controls. The remaining three categories were not statistically significant. The numbers indicate mean percentage values. The bars indicate standard deviations. Mann Whitney two-tailed test was performed to calculate *P*-values (*** - *P*<0.001). D) The profile plots show the distribution of the poly(dA) sequence (AAAA) (dashed line), the predicted G4 structures (solid line) and the nucleosome distribution of the replicate GSM2407365 from BSF cells [49] (blue dotted line) around centred origins and intergenic shuffled controls (± 2 kb). The G4s were predicted by the G4Hunter application [50] in the Tb Lister 427 reference genome. 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).

We then quantified the number of origins without or with HNO regions on one or both sides of the centre, and the presence or absence of LNO in the centre and compared them to the intergenic shuffled controls (Methods). We performed this quantification for the merged set of origins (Figure 6B), but also for the mapped origins in PCF and BSF cells against four replicates of the PCF and four replicates of the BSF MNase-seq datasets [49] (Supplementary Figure 6B). Six categories of different combinations of HNO and LNO were analysed, and quantification showed that there were statistically significant differences between origins and corresponding shuffled controls when PCF and BSF origins were merged (Figure 6B) and when PCF and BSF origins were analysed separately (Supplementary Figure 6B). We then analysed each of the six categories individually to determine which had changed significantly between the origins and the shuffled controls. It was found that three categories were significantly different between origins and shuffled controls, while three categories were not (Figure 6C). The category two HNO flanking LNO was significantly enriched (twice more), whereas the category no HNO with LNO and the category no HNO no LNO were significantly depleted in the origins compared to the shuffled controls (Figure 6C). Therefore, the active origins had significantly more LNO flanked by two HNO, at the expense of the no HNO with LNO and the no HNO no LNO categories.

Further investigation was undertaken to determine the spatial relationship between poly(dA)/G4 enriched sequences and nucleosome distribution. To achieve this, strand separated poly(dA) enriched sequences and predicted G4s were overlapped with the MNase-seq data [49] (Methods). The plot obtained after the intersections revealed that the LNO region was positioned between the plus and minus poly(dA)/G4 summits (Figure 6D). Additionally, the HNO peaks were found to exhibit overlap with the poly(dA) and G4 peaks and to surround the origins (Figure 6D). 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 6D).

The origins colocalized 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 [51]. 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 [52, 53]. A recent study employed DRIP-seq to map R-loops in *T. brucei* BSF cells, revealing their enrichment within intergenic regions, between coding sequences of polycistronic transcription units, and coinciding with sites of polyadenylation and nucleosome depletion [54]. 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 [54] was conducted (Methods). A significant accumulation of R-loops or RNA:DNA hybrids was identified around the origins, with two distinct summits located close to the origin centre (Figure 7A). An R-loop enrichment was also evident in the vicinity of the centre of the intergenic shuffled controls (Figure 7A), which is consistent with previous reports indicating a high prevalence of R-loops in intergenic regions [54]. 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 controls (Figure 7A). Next, we quantified, the number of origins and shuffled controls that overlapped with R-loops [54]. 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 $2.87E^{-142}$), suggesting a link between RNA:DNA hybrids formation and origin activity. We also checked the distribution of R-loops detected in BSF cells around PCF and BSF origins separately and found that the distribution of RNA:DNA hybrids was similar around BSF origins and equivalent to the merged set of origins (Supplementary Figure 7A). The distribution of RNA:DNA hybrids was also enriched around the centres of PCF origins with one of the upper

summits less pronounced (Supplementary Figure 7A). Quantification of R-loops overlapping origins was performed and it was shown that 86% of PCF origins and 93% of BSF origins overlapped with R-loops (Supplementary Figure 7B). It is worth noting that these intersections also revealed that only a small proportion of the previously identified R-loops [54] overlapped with the merged origins (1.7%), PCF origins (0.7%), or BSF origins (1.3%).

Subsequently, the predicted G4 structures and poly(dA) enriched sequences from the plus and minus strands were superimposed with the DRIP-seq dataset [54] 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 distinct 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 7C).

Comparison of the origins identified using stranded SNS-seq with previously published MFA-seq and TbORC1/CDC6 binding sites datasets

We compared the replication origins identified by stranded SNS-seq with the data obtained by Marker frequency analysis coupled with deep sequencing (MFA-seq) [42, 43]. MFA-seq compares read coverage between replicating and non-replicating cells and identifies regions with increased coverage in S phase as replication origins. In *T. brucei*, MFA-seq identified broad regions of a few hundred kilobases [42, 43], whereas the origins detected by stranded SNS-seq are about 150 bp long. We used six MFA-seq datasets [42, 43] (Methods) and identified the previously published 47 MFA-seq peaks [43]. We then intersected the MFA-seq datasets with the SNS-seq dataset (Supplementary Figure 8A) to assess the extent of the overlap (Supplementary Figure 8B). We found 28-42% stranded SNS-seq origins overlapped with early and 43-55% overlapped with late S-phase MFA-seq replicated regions (Supplementary Figure 8B).

Several proteins have been identified as potential subunits of the origin recognition complex (ORC) in *T. brucei*. One of these is TbORC1/CDC6, which replaced the yeast protein Cdc6, but not Orc1, in a phenotypic complementation assay [55]. We intersected the origins detected by stranded SNS-seq with the previously detected binding sites of the TbORC1/CDC6-12Myc tagged protein [42] (Methods). For this intersection, we used the only publicly available data from TriTrypDB describing 350 ORC1/CDC6 binding sites, which differs from the 953 TbORC1/CDC6 binding sites described previously [42] (Methods). The intersection revealed that TbORC1/CDC6 levels are lowest near the centre of origins (Supplementary Figure 8C). An overlap of 4.2% was obtained between origins and TbORC1/CDC6 binding sites within a ±2 kb window, 6.2% within a ±3 kb window and 19.9 % in ±10 kb window.

Spatial coordination between the activity of the origin and transcription

DNA replication in trypanosomatids occurs in a particularly challenging environment, as most of their genomes are constitutively transcribed. Arrays of genes oriented in the same direction are transcribed by RNA polymerase II into polycistronic transcription units (PTUs), which can contain hundreds of genes (reviewed in [40]). The primary transcript is processed into individual mRNAs by trans-splicing of a capped ‘spliced leader’ at the splice acceptor site (SAS) and by polyadenylation at the polyadenylation sites (PAS) (reviewed in [40]). To gain insight into the relationship between origins and SAS and PAS, we intersected our dataset with the unified SAS and PAS datasets for the core chromosomes (Methods). The intersection of origins with SAS showed that the centre of origins was depleted of SAS, while some origins had SAS flanking the centre and the SAS signal was enriched ±200-500 nt from the centre of origins (Figure 7D, left panel). Only 15% of the origins overlapped with SAS compared to 24% of the shuffled controls, and this difference was statistically significant (Figure 7D, right panel). The intersection of the origins with PAS revealed an accumulation of PAS near the origin centres. A similar, though less pronounced, accumulation

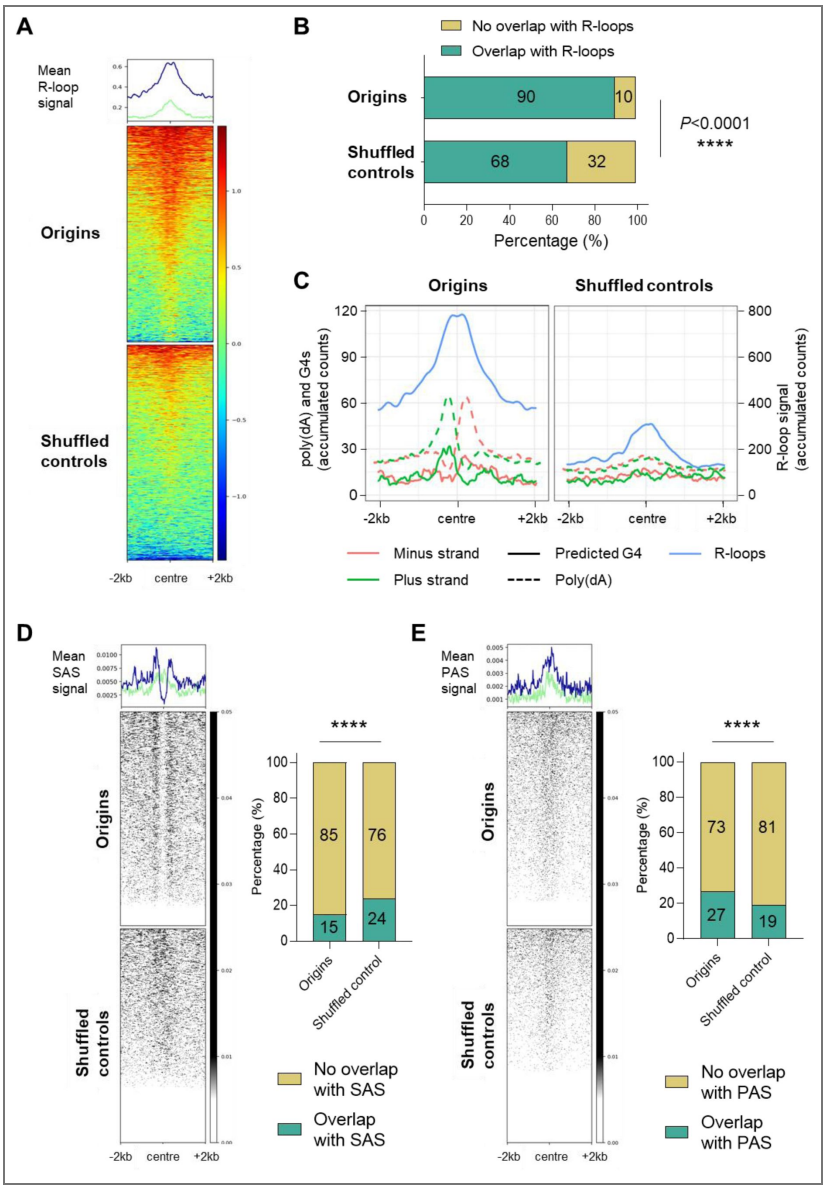

Figure 7. The distribution of R-loops, splice acceptor sites and polyadenylation sites around the origins.

A) The profile plots and heatmaps show the distribution of R-loops [54] around the centred origins (blue line) and the intergenic shuffled controls (green line) (± 2 kb). Mean DRIP-seq signal presents average DRIP-seq signal [54] per 20 bp window. B) The proportion of origins and shuffled controls that overlap with R-loops (intersection without window). The *P*-values were calculated using Fisher's exact (two-sided) 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 [54] (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 [50] in the Tb Lister 427-2018 reference genome. 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). D) Left panel: the profile plots and heatmaps show the distribution of transcription splice acceptor sites (SAS) around the centred origins (blue line) and the intergenic shuffled controls (green line) (± 2 kb). Intersection was performed in the Tbb TREU927 reference genome for the 11 megabase chromosomes. Right panel: the proportion of origins and shuffled controls that overlap with SAS (intersection without window). The *P*-value was calculated using Fisher's exact (two-sided) test, which involved a comparison of the numbers of indicated categories for each pair of datasets (**** - *P*<0.0001). E) Left panel: the profile plots and heatmaps show the distribution of transcription polyadenylation sites (PAS) around the centred origins (blue line) and the intergenic shuffled controls (green line) (± 2 kb). Intersection was performed as in Figure 7D. Right panel: the proportion of origins and shuffled controls that overlap with PAS. The *P*-values were calculated using Fisher's exact (two-sided) test, which involved a comparison of the numbers of indicated categories for each pair of datasets (**** - *P*<0.0001).

was also observed in the shuffled controls (Figure 7E, left panel). Quantification showed that 27% of origins overlapped with PAS compared to 19% of shuffled controls and this difference was statistically significant (Figure 7E, right panel).

PTUs are separated by strand-switching regions (SSRs), which are classified as either divergent (dSSRs) or convergent (cSSRs). dSSRs and cSSRs serve as transcription start and termination regions, respectively (reviewed in [40]). Our results showed that the origins are predominantly located in the intergenic regions (Figure 2C). Only 0.4% of origins are located within the dSSRs and 0.3% of origins are located within the cSSRs. Therefore, most origins are localized in actively transcribed regions, which could lead to collisions between DNA replication and the transcription machinery. This spatial coincidence implies that transcription and replication must occur in a highly ordered and cooperative manner in *T. brucei*.

Discussion

We used stranded SNS-seq to identify active origins in the unicellular eukaryote, *T. brucei*. Our methodology is comparable to previously published origin-derived single-stranded DNA sequencing (Ori-SSDS) [17] but includes few differences. The stranded library preparation process was different in these two approaches, and the SNS-depleted control was included in the origin mapping process 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 of an active origin. In addition, this method maps the origins between two divergent SNS peaks, allowing the determination of the centre of origins or replication initiation sites, which was essential for the discovery of the positions and orientations of the elements involved in the topology of origins. Origins mapped between two divergent SNS peaks showed the origin-specific nucleotide distribution around the centre (Figure 4A, Supplementary Figure 9A). On the other hand, if we apply conventional SNS-seq origin mapping to our data and map origins in the middle of peaks without knowing their orientation, the origin-specific nucleotide distribution is lost (Supplementary Figure 9B). Therefore, the primary benefits of our approach include the ability to centre origins between two divergent SNS peaks and to eliminate false positive peaks.

Using this stringent origin selection procedure, we have selected and analysed a subset of all origins. Some weak or inefficient origins may be missed due to low activity, resulting in low read coverage. It is therefore very likely that the number of origins of this parasite is higher than the number reported here. We also examined the origins separately in the two life stages demonstrating that the origin structures in the PCF and BSF cells were equivalent and justifying the merging of these two sets of origins into one. Under this stringent origin mapping condition, we found that 45% of PCF origins were shared with BSF cells (Supplementary Figure 2C). However, we decided not to compare the PCF and BSF sets of origins in more detail because comparing non-exhaustively mapped origins in these two life cycle stages could be misleading.

Our analysis revealed no consensus sequence of origins, which is consistent with the findings of most other studies of eukaryotic origins, including trypanosomatids [22, 27, 42, 43]. Nevertheless, the genomic sites that function as origins are not random sequences. Thanks to the stranded SNS-seq approach, we were able to determine the positions and strand-specificity of some genetic elements and to present a comprehensive model of an origin (Figure 8). The results of our study showed that 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's centre (Figure 4B and Figure 8). The G4s 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). The centres of origin were identified as LNO regions, which had an average length of 200–300 nt and were flanked by HNO regions located from ±150 to ±500 bp from the centre (Figure 6D, Figure 8). RNA:DNA hybrid enrichment revealed two distinct summits,

one upstream and one downstream of the centre, with a range spanning ±200 nucleotides (Figure 7A and Figure 8). Our model suggests that the leading strands of DNA replication, or SNS molecules, are synthesised using the poly(dA)/G4-rich strands as templates.

Our analysis indicated that G-rich sequences near origins tend to form G4 structures (Figure 5A-C, Supplementary Figure 5B-C). 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 propose 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 [21, 56]. 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.

Our study disclosed for the first time a strong link between RNA:DNA hybrids formation and DNA replication initiation. We think that high overlap between two distinct techniques – DRIP-seq and stranded SNS-seq – provides compelling validation of stranded SNS-seq origin mapping. We suggest that the observed enrichment of RNA:DNA hybrids near the origins may represent the DNA replication initiation sites. The S9.6 antibody used for immunoprecipitation in DRIP-seq experiments exhibits a high affinity for RNA:DNA hybrids longer than 6 bp [57]. The RNA:DNA hybrids are formed at the replication initiation sites by RNA priming of the SNS and Okazaki fragments and S9.6 antibody can immunoprecipitate RNA:DNA hybrids formed between these molecules and genomic DNA. However, Okazaki fragments are not exclusive markers of origins as they form RNA:DNA hybrids throughout the genome, whereas RNA:DNA hybrids formed by RNA-primed SNS are specific markers of origins. Therefore, it can be suggested that the RNA:DNA hybrid summits observed next to origins were most likely the result of an enrichment of RNA-primed SNS molecules.

The elements presented in the model were not found in all origins, as revealed by the quantification in this study. This discrepancy may be caused by the fact that the datasets used for the intersections were obtained from four different strains, rather than from just one. Different strains are likely to activate different sets of origins, which could introduce bias during the intersection process. Another possible source of bias may be the varying quality of the genome assemblies, given that this study used three different reference genomes (Methods). This is clearly demonstrated by the detection of different numbers of origins in different reference genomes using the same stranded SNS-seq data (Methods). An alternative hypothesis that could explain the absence of certain elements in some origins is that these structures form very transiently, only during the moment of origin firing. Therefore, it can be hypothesised that the presence of these elements, in addition to the chromatin state and nuclear organisation of origins, defines the moment of origin activation.

Different origin mapping methods often demonstrate limited overlap due to biases related to the chosen methodology [58, 59]. A comparison of MFA-seq and stranded SNS-seq origins showed that they partially overlap in *T. brucei* (Supplementary Figure 8B). It is worth noting that a similar result was found in another trypanosomatid, *Leishmania major*, when conventional SNS-seq and MFA-seq were analysed [22]. As it is not possible to determine how many individual DNA replication origins are present within the broad MFA-seq replicated regions, it is impossible to assess whether there is any direct overlap with the SNS-seq origins. Other factors could also explain the variations between stranded SNS-seq and MFA-seq. Firstly, MFA-seq relies on cell sorting, which has the potential to introduce bias into the overall origin mapping process. In contrast, stranded SNS-seq is performed on asynchronous cell populations, enabling the identification of origins active throughout S phase. Secondly, the hypothesis that MFA-seq does not detect all active origins is supported by combing data, which revealed a much smaller IOD (~200 kbp) at the single-molecular level than that suggested by MFA-seq at the population level (~600 kbp) [42]. Thirdly, as MFA-seq calculates the ratio of read coverage between the S and G2 phases, it may lack the

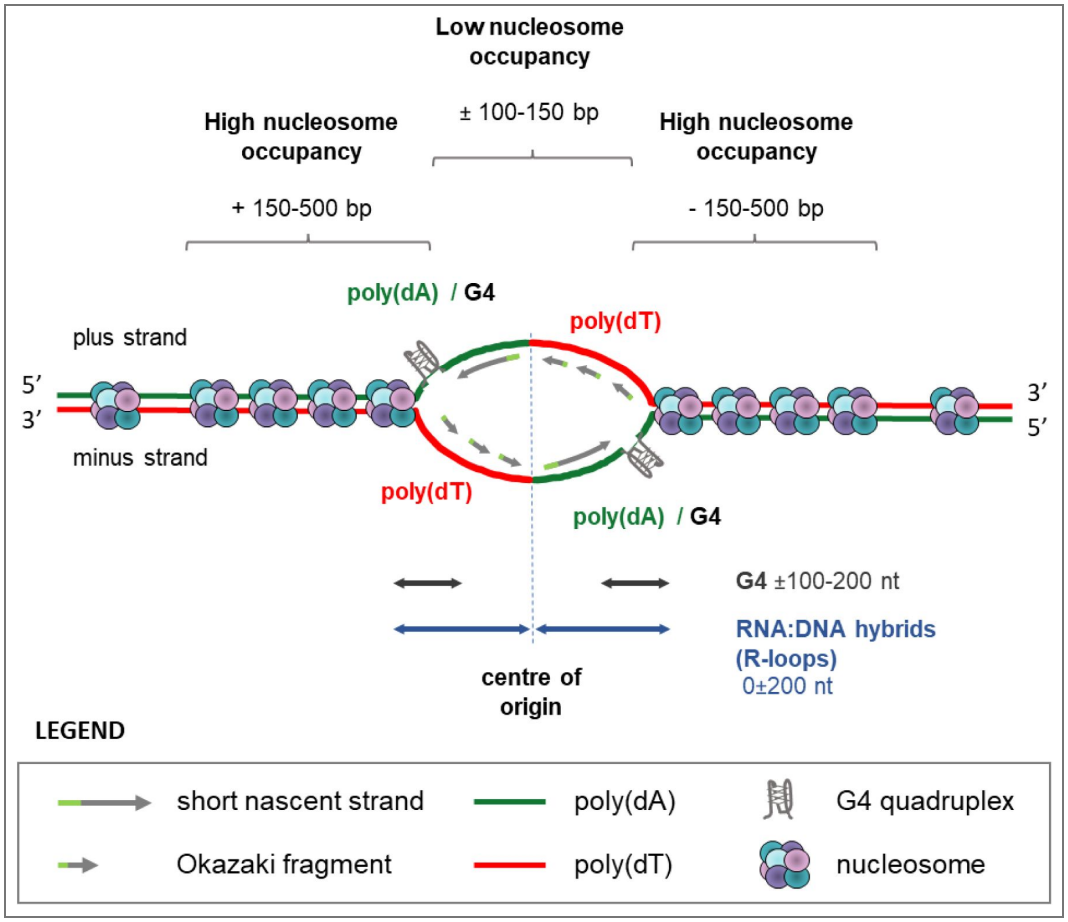

Figure 8. A proposed model of origin with the position of different genetic elements and nucleosome occupancy.

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, LNO 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.

sensitivity required to detect origins used by fewer than 25% of cells in the population [60]. Similarly, stringent origin detection using stranded SNS-seq identified a subset of origins. Therefore, comparing two incomplete datasets has its limitations and may not be relevant.

Validation of origins is generally a difficult task, particularly in trypanosomatids, where proteins involved in the initiation of DNA replication are difficult to determine. Few proteins have been described as potential ORC subunits (reviewed in [61]), and none of them have been shown to be a specific marker that indicates the origins. We found a minority of TbORC1/CDC6 binding sites near origins detected by SNS-seq and this was comparable with the previously reported ~4.4% overlap of this protein with MFA-seq data (reviewed in [61]). Comparable results were also obtained in *T. cruzi* [27]. DNAscent nanopore sequencing revealed 4,287 origins in *T. cruzi*. Of these, 3,782 (88.2%) did not match the 851 TcORC1/CDC6-Ty1 binding sites within a ±3 kb window [27]. Therefore, only 11.7% of origins of *T. cruzi* overlapped with TcORC1/CDC6-Ty1 binding sites within a ±3 kb window, a proportion similar to what we observed in *T. brucei* (6.2%). It is also worth noting that origins detected by conventional SNS-seq in *Plasmodium falciparum* revealed 63% of overlap with origins detected by nanopore sequencing and 13.6% with $PfORC_{1\text{-}2}$ proteins [36]. Another group found no significant correlation between origins detected by DNAscent nanopore sequencing and PfORC1 ChIP sites within the same organism[62].

It is important to note that the topology of origins of *T. brucei* is very similar to that previously described for other 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 [13]. The same position and orientation of predicted G4 motifs, peaking at 240-300 base pairs from replication start sites, was previously identified in two Drosophila cell lines using conventional SNS-seq [21], as was found here for *T. brucei*. Furthermore, the organisation of nucleosomes around the origins of *T. brucei* is identical to that observed in yeast and higher eukaryotes [12, 28–32, 63, 64]. In the case of *L. major* the origins identified using conventional SNS-seq were also located in genomic regions with lower nucleosome occupancy [22]. In addition, a significant correlation between *L. major* origins and predicted G4 motifs was observed, with 74% of origins containing at least one such motif [22]. These similarities suggests that the structure of origins may be a conserved feature in all eukaryotes. However, further studies are needed to gain a deeper insight into the evolution of DNA replication origins.

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 [65]. 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.* [66], 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

SNS were isolated and purified following a previously published protocol [67] with few modifications. Between 2 - 2.5×10^9 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 (3-5×10^6 and 7-9×10^5 $cells.mL^{-1}$ 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 (10 mM Tris-Cl, 50 mM EDTA, 20 mM NaCl, pH 7.0) 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 +4°C. 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 10A). 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, Cat Nb 11119915001) 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 (95°C/10 min) and simultaneously on samples and controls. After denaturation, samples and controls were immediately placed on ice. The enzymatic reactions were assembled on ice, and then the enzymatic treatment with T4 PNK and λ-exonuclease was done at 37°C. After enzymatic treatment and prior to library preparation, residual RNA was removed with 5 µg RNase, DNase-free (Roche Cat Nb 11119915001), isolated DNA was sonicated with Bioruptor Pico sonication device (Diagenode) and purified with magnetic beads (CleanNGS, Proteigen). The quantity of purified 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 [20, 67]. 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 [20]. 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 10B). RNA molecules treated under the same conditions were preserved (Supplementary Figure 10C), suggesting that SNS would also be preserved. The quantity of DNA molecules 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

Single-stranded SNS libraries were prepared using the Accel-NGS 1S Plus DNA Library Kit and 1S Plus Dual Indexing Kit for Illumina (Swift Biosciences), following the manufacturer's protocol. 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 (2×150)

The libraries were validated for quality and quantity using 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 (*RRID:SCR_015058*, v2.20.0.422).

SNS-seq data analysis

Quality control of the FASTQ files was performed using FastQC (*RRID:SCR_014583*, v0.11.9) (*https://www.bioinformatics.babraham.ac.uk/projects/fastqc/*). The sequencing reads were tail trimmed to remove the 3'end specific tail introduced during library preparation using cutadapt [68] 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 427-2018 reference genome release 55 from TriTrypDB (*RRID:SCR_007043*) (*https://tritrypdb.org/common/downloads/release-55/TbruceiLister427_2018*, last accessed June 24th, 2025) with bowtie2 (*RRID:SCR_016368*, v2.4.10) [69]. The alignment output bam files were converted, sorted, and indexed using samtools[70]. Duplicated reads were removed using picard (*RRID:SCR_006525*, v2.23.5, "Picard Toolkit." 2019. Broad Institute, GitHub Repository. *https://broadinstitute.github.io/picard/*, last accessed June 24th, 2025; Broad Institute). Mapped deduplicated reads pairs were then separated with samtools (*RRID:SCR_002105*, v1.13) *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 with the SDC as control.

Origin selection

Origins were selected after the peak calling step, during the peak filtering step. The called peaks were filtered using the bedtools suite (*RRID:SCR_006646*, v.2.30.0) [71] 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 two divergent minus and plus peak pair. The distances between the start points of the two divergent leading strands are not known. We conducted tests on several inner border-to-border distances, ranging from 0.5 to 2.5 kb. Higher numbers of origins were found with larger distances. However, we finally applied a maximum distance of 0.5 kb, presuming that this corresponds more closely to the physiological distance between two divergent SNS molecules. The third criterion was introduced to prevent complete overlap of two divergent peaks, as this can only occur if the two SNSs 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** Integrative genomics viewer (IGV) (*RRID:SCR_011793*) [72] was used. Violin plots were generated with the GraphPad Prism 8.3.0 software (*RRID:SCR_002798*, 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 samples and control and BSF samples and control, respectively, were merged using Samtools *merge*. The PCF combined sample was sub-samples with Samtools *view* (--subsample 0.69--subsample-seed [0-14]) and the merged control with Samtools *view* (--subsample 0.57--subsample-seed [0-14]) 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 (*RRID:SCR_016366*, v3.5.1) [73] and the ggplot2 R package was used. The open source ggplot2 R package (*RRID:SCR_014601*), freely available from the Comprehensive R Archive Network (CRAN) repository (*http://www.r-project.org*, last accessed June 24th, 2025) was used to generate composite plots. 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.

Motive enrichments were tested over the origin centres extended 250 bp up and downstream with the MEME suite motive discovery (*RRID:SCR_001783*) [74] as well as with homer (*RRID:SCR_010881*) *findMotifs* [75]. A multiple sequence alignment and the analysis of nucleotide patterns were built with the WebLogo (*RRID:SCR_010236*) online application [46].

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 June 24th, 2025).

Comparison of origins mapped by stranded SNS-seq with previously published datasets

Handling of datasets mapped to different reference genomes

The stranded SNS-seq data, that we mapped to the Tb Lister 427-2018 genome, were compared with various published datasets (SAS and PAS (TriTrypDB), MNase-seq [49], DRIP-seq [54], G4-seq [48], MFA-seq [42] and Orc1/CDC6 ChIP-on-Chip [42]) that were analysed using different references genomes (Tb Lister 427 or TREU927). To enable direct comparison, we either reanalysed the published datasets against the Tb Lister 427-2018 genome or remapped the stranded SNS-seq data to match the respective reference genomes of the compared datasets. The mapping of origins in the different reference genomes produced different numbers of origins, in Tb Lister 427 1059 PCF and 1466 BSF origins and in total 2081 origins were detected, in TREU927 976 PCF and 1466 BSF origins were detected and in total 1963 origins were detected. Specific details on the chosen approach are given in the relevant paragraphs below.

Splice leader acceptor sites (SAS) and polyadenylation sites (PAS) were obtained from TriTrypDB. The tracks 'unified spliced leader addition sites' and the 'unified poly A sites' were downloaded from the genome browser view for the 11 Megabase chromosomes of the Tbb TREU927 reference genome in bed file format. We compared these datasets to SNS-seq dataset analysed as described above in the Tbb TREU927 reference genome (*https://tritrypdb.org/common/downloads/release-62/TbruceiTREU927/*).

MNase-seq

The MNase-seq data [49] was downloaded as bigwig files from GEO (accession number GSE90593). The data was published using the Tb Lister 427 reference genome. To be able to compare our SNS-seq dataset with this data we reanalysed the SNS-seq sequencing reads as described above in the Tb Lister 427 reference genome (*https://tritrypdb.org/common/downloads/release-62/TbruceiLister427/* last accessed June 24th, 2025). To quantify the MNase-seq signature we observed around the origin centres ("M" pattern), we extracted the average MNase-seq signal (normalised dyad values [49]) from the regions 400-200 nt upstream, 200-400 nt downstream and 75nt up- and downstream of the origin centre with the UCSC utilities *bigWigAverageOverBed* command (*https://github.com/ucscGenomeBrowser/kent/tree/master/src/utils* last accessed June 24th, 2025). Regions with negative values representing tandem repeats and N-rich sequences [49] were discarded. LNO regions were considered as present when the average MNase-seq value was below 0.23 normalised dyad value and the HNO regions were considered present when the averaged normalised dyad value for the region was above this value. This value corresponds to the average value for background regions (2000-1000 bp upstream of the origin centres and 1000-2000 bp downstream of origin centres).

Drip-seq

The Drip-seq dataset was downloaded from ENA (accession number PRJEB21868) [54] as fastq files. The re-analysis was done as described [54]. The published DRIP-seq analysis was performed in Tb Lister 427 reference genome. We re-analysed the data in the Tb Lister 427 and the Tb Lister 427-2018 reference genomes. The intersection of DRIP-seq data with the stranded SNS-seq origins was done in the Tb Lister 427-2018 reference genome.

G4-seq

The G4-seq [48] dataset was downloaded from GEO (accession number GSE110582) as bed and bedgraph files separated for the plus and minus strand. G4-seq was performed on *T. brucei* EATR01125 strain and analysed in the Tbb TREU927 reference genome [48] to be able to directly compare the data with the SNS-seq dataset we analysed the SNS-deq data in the Tbb TREU927 reference genome.

G4 prediction

In the *T. brucei* genome were done with G4Hunter [50] with the following settings, window size (-w 25) and a threshold of 1.57. Prediction of G4 structures in the Tb 427-2018 reference genome and in the Tb Lister 427 reference genome were performed, to be able to compare G4 structures to MNase-seq (in Tb Lister 427 reference genome) and DRIP-seq (in Tb Lister 427-2018 reference genome) data as G4-seq output files were available only for the Tbb TREU927 reference genome [48].

MFA-seq

We intersected six MFA-seq datasets. Two were obtained directly from the authors[42] and four were downloaded from ENA (accession number PRJEB11437) [43], all in fastq file format. The analysis of the fastq files was performed as described previously [42, 43]. The original published MFA-seq analysis was performed in *T. brucei* Tbb TREU927 reference genome. We re-analyzed the data in the Tbb TREU927 reference genome and in the Tb Lister 427-2018 reference genomes to be able to directly compare them to the stranded SNS-seq data.

ORC1/CDC6 ChIP-on-chip

The Orc1/CDC6 ChIP-on-Chip dataset [42] was downloaded from TriTrypDB (as ratios_peaks) for each element in the *T. brucei* Tbb TREU927 reference genome in form of bed files. The dataset available on TriTrypDB differs from the one published in the manuscript of Tiengwe et al. [42] but the originally published dataset is not available anymore (personal communication Richard McCulloch). The TriTrypDB ORC1/CDC6 dataset contains 350 binding sites whereas the dataset published originally contained 953 binding sites [42]. To be able to directly compare them to the SNS-seq we analysed the SNS-seq data in the Tbb TREU927 reference genome.

DNA molecular combing

Asynchronous cell populations of *T. brucei* were grown to the density of 4×10^6 and 7×10^5 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 1×10^8 cells were blocked in 1% low melting agarose plugs.

DNA molecular combing was done as described previously [41]. 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 molecules were combed and regularly stretched (2 kb/µm) on silanized coverslips provided by Montpellier DNA Combing Facility. Linearized DNA molecules 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 X-100. Immuno-detection was done with antibodies diluted in 1 × PBS/ 1% BSA/ 0.1% Triton X-100 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, Cat. No. MAB3034, Merck Milipore), the mouse anti-BrdU antibody (1/20 dilution, clone B44, Cat. No. 347580, Becton Dickinson) and rat anti-BrdU antibody (1/20 dilution, clone BU1/75, Cat. No. ABC117-7513, Abcys). The following secondary antibodies were used: Alexa488 Goat anti Rat IgG (1/50 dilution, Cat. No. A-11006, Invitrogen), Alexa 546 Goat anti Mouse IgG1 (1/50 dilution, Cat. No. A-21123, Invitrogen) and Alexa 647 goat anti-mouse IgG2a (1/100 dilution, Cat. No. A-21241, Invitrogen). Coverslips were mounted with 15 µL of Prolong Gold Antifade (Cat. No. P10144, Thermo Fisher Scientific) 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) (*RRID:SCR_002368*).

Measurements and statistical analysis of combed DNA molecules

The speed of replication forks was estimated on individual forks, presented as IdU track followed by a CldU track. Only intact (non-fragmented) 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 [41]. The statistical comparisons of different datasets were assessed by the non-parametric Mann–Whitney two-tailed tests.

Data availability

The SNS-seq data (FASTQ files) and the BED files containing the positions of the origins of the merged and separated PCF and BSF samples 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* and *https://zenodo.org/records/14289282*.

Acknowledgements

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 [76]. 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.

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).

Additional files

Supplementary Table 1-2 and Supplementary Figure 1-10

Additional information

Funding

Funder	Grant reference number
Agence Nationale de la Recherche	ANR-11-LABX-0024-01 "PARAFRAP"
Agence Nationale de la Recherche	ANR-10-INBS-09
Agence Nationale de la Recherche	ANR-11-INBS-0013
Fondation pour la Recherche Médicale	ING20160435527

Author ORCID iDs

Slavica Stanojcic: http://orcid.org/0000-0003-4505-7660
Bridlin Barckmann: http://orcid.org/0000-0001-8354-668X
Pieter Monsieurs: http://orcid.org/0000-0003-2214-6652
Yvon Sterkers: http://orcid.org/0000-0002-5623-5664

References

1. **Lee C.S.K.,** Weiβ M., Hamperl S. (2023) Where and when to start: Regulating DNA replication origin activity in eukaryotic genomes. *Nucleus* **14**
2. **Hulke M.L.,** Massey D.J., Koren A (2020) Genomic methods for measuring DNA replication dynamics. *Chromosome Res* **28**:49-67
3. **Leonard A.C.,** Méchali M. (2013) DNA replication origins. *Cold Spring Harb Perspect Biol* **5**:a010116

4. **Newlon C.S.,** Theis J.F (1993) The structure and function of yeast ARS elements. *Curr Opin Genet Dev* **3**:752-758
5. **Okuno Y.,** Satoh H., Sekiguchi M., Masukata H (1999) Clustered adenine/thymine stretches are essential for function of a fission yeast replication origin. *Mol Cell Biol* **19**:6699-6709
6. **Kong D.,** 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
7. **MacAlpine D.M.,** Rodríguez H.K., Bell S.P (2004) Coordination of replication and transcription along a Drosophila chromosome. *Genes Dev* **18**:3094-3105
8. **Balasov M.,** Huijbregts R.P., 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
9. **Stanojcic S.,** Lemaitre J.M., Brodolin K., Danis E., Mechali M (2008) In Xenopus egg extracts, DNA replication initiates preferentially at or near asymmetric AT sequences. *Mol Cell Biol* **28**:5265-5274
10. **Segurado M.,** de Luis A., Antequera F (2003) Genome-wide distribution of DNA replication origins at A+T-rich islands in Schizosaccharomyces pombe. *EMBO Rep* **4**:1048-1053
11. **Mojardín L.,** Vázquez E., Antequera F (2013) Specification of DNA replication origins and genomic base composition in fission yeasts. *J Mol Biol* **425**:4706-4713
12. **Eaton M.L.,** Galani K., Kang S., Bell S.P., MacAlpine D.M (2010) Conserved nucleosome positioning defines replication origins. *Genes Dev* **24**:748-753
13. **Tubbs A.,** et al. (2018) Dual Roles of Poly(dA:dT) Tracts in Replication Initiation and Fork Collapse. *Cell* **174**:1127-1142.e1119
14. **Delgado S.,** Gómez M., Bird A., Antequera F (1998) Initiation of DNA replication at CpG islands in mammalian chromosomes. *EMBO J* **17**:2426-2435
15. **Cadoret J.C.,** et al. (2008) Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc Natl Acad Sci U S A* **105**:15837-15842
16. **Cayrou C.,** 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
17. **Pratto F.,** et al. (2021) Meiotic recombination mirrors patterns of germline replication in mice and humans. *Cell* **184**:4251-4267.e4220
18. **Besnard E.,** et al. (2012) Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat Struct Mol Biol* **19**:837-844
19. **Valton A.L.,** et al. (2014) G4 motifs affect origin positioning and efficiency in two vertebrate replicators. *EMBO J* **33**:732-746
20. **Cayrou C.,** et al. (2015) The chromatin environment shapes DNA replication origin organization and defines origin classes. *Genome Res* **25**:1873-1885
21. **Comoglio F.,** et al. (2015) High-resolution profiling of Drosophila replication start sites reveals a DNA shape and chromatin signature of metazoan origins. *Cell Rep* **11**:821-834
22. **Lombraña R.,** et al. (2016) Transcriptionally Driven DNA Replication Program of the Human Parasite Leishmania major. *Cell Rep* **16**:1774-1786
23. **Langley A.R.,** Gräf S., Smith J.C., 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
24. **Sequeira-Mendes J.,** et al. (2019) Differences in firing efficiency, chromatin, and transcription underlie the developmental plasticity of the. *Genome Res* **29**:784-797
25. **Akerman I.,** et al. (2020) A predictable conserved DNA base composition signature defines human core DNA replication origins. *Nat Commun* **11**:4826
26. **Prorok P.,** et al. (2019) Involvement of G-quadruplex regions in mammalian replication origin activity. *Nat Commun* **10**:3274

27. **Vitarelli M.d.O.,** et al. (2024) Integrating high-throughput analysis to create an atlas of replication origins in Trypanosoma cruzi in the context of genome structure and variability. *mBio* **15**:e00319-00324
28. **Lipford J.R.,** Bell S.P (2001) Nucleosomes positioned by ORC facilitate the initiation of DNA replication. *Mol Cell* **7**:21-30
29. **Berbenetz N.M.,** Nislow C., Brown G.W (2010) Diversity of eukaryotic DNA replication origins revealed by genome-wide analysis of chromatin structure. *PLoS Genet* **6**:e1001092
30. **Foss E.J.,** et al. (2024) Identification of 1600 replication origins in *S. cerevisiae*. *eLife* **12**:RP88087
31. **Poulet-Benedetti J.,** et al. (2023) Dimeric G-quadruplex motifs-induced NFRs determine strong replication origins in vertebrates. *Nat Commun* **14**:4843
32. **Jansen A.,** Verstrepen K.J (2011) Nucleosome positioning in Saccharomyces cerevisiae. *Microbiol Mol Biol Rev* **75**:301-320
33. **Nasheuer H.P.,** Meaney A.M (2024) Starting DNA Synthesis: Initiation Processes during the Replication of Chromosomal DNA in Humans. *Genes* **15**
34. **Bielinsky A.K.,** Gerbi S.A (1998) Discrete start sites for DNA synthesis in the yeast ARS1 origin. *Science* **279**:95-98
35. **Foulk M.S.,** Urban J.M., Casella C., 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
36. **Castellano C.M.,** et al. (2024) The genetic landscape of origins of replication in P. falciparum. *Nucleic Acids Res* **52**:660-676
37. **Senkevich T.G.,** et al. (2015) Mapping vaccinia virus DNA replication origins at nucleotide level by deep sequencing. *Proc Natl Acad Sci U S A* **112**:10908-10913
38. **Fenn K.,** Matthews K.R (2007) The cell biology of Trypanosoma brucei differentiation. *Curr Opin Microbiol* **10**:539-546
39. **Baldauf S.L** (2003) The deep roots of eukaryotes. *Science* **300**:1703-1706
40. **Clayton C** (2019) Regulation of gene expression in trypanosomatids: living with polycistronic transcription. *Open Biol* **9**
41. **Stanojcic S.,** et al. (2016) Single-molecule analysis of DNA replication reveals novel features in the divergent eukaryotes *Leishmania* and *Trypanosoma brucei* versus mammalian cells. *Sci Rep* **6**
42. **Tiengwe C.,** 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
43. **Devlin R.,** et al. (2016) Mapping replication dynamics in Trypanosoma brucei reveals a link with telomere transcription and antigenic variation. *eLife* **5**:e12765 https://doi.org/10.7554/eLife.12765
44. **Fragkos M.,** Ganier O., Coulombe P., Mechali M (2015) DNA replication origin activation in space and time. *Nat Rev Mol Cell Biol* **16**:360-374
45. **Techer H.,** et al. (2013) Replication dynamics: biases and robustness of DNA fiber analysis. *J Mol Biol* **425**:4845-4855
46. **Crooks G.E.,** Hon G., Chandonia J.M., Brenner S.E (2004) WebLogo: a sequence logo generator. *Genome Res* **14**:1188-1190
47. **Valton A.L.,** Prioleau M.N (2016) G-Quadruplexes in DNA Replication: A Problem or a Necessity?. *Trends Genet* **32**:697-706
48. **Marsico G.,** et al. (2019) Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic Acids Res* **47**:3862-3874
49. **Maree J.P.,** Povelones M.L., Clark D.J., Rudenko G., Patterton H.G (2017) Well-positioned nucleosomes punctuate polycistronic pol II transcription units and flank silent. *Epigenetics Chromatin* **10**

50. **Brázda V.,** et al. (2019) G4Hunter web application: a web server for G-quadruplex prediction. *Bioinformatics* **35**:3493-3495
51. **Sollier J.,** Cimprich K.A (2015) Breaking bad: R-loops and genome integrity. *Trends Cell Biol* **25**:514-522
52. **Boguslawski S.J.,** et al. (1986) Characterization of monoclonal antibody to DNA.RNA and its application to immunodetection of hybrids. *J Immunol Methods* **89**:123-130
53. **Wahba L.,** Costantino L., Tan F.J., Zimmer A., Koshland D (2016) S1-DRIP-seq identifies high expression and polyA tracts as major contributors to R-loop formation. *Genes Dev* **30**:1327-1338
54. **Briggs E.,** Hamilton G., Crouch K., Lapsley C., 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
55. **Godoy P.D.,** et al. (2009) Trypanosome prereplication machinery contains a single functional orc1/cdc6 protein, which is typical of archaea. *Eukaryot Cell* **8**:1592-1603
56. **Zerbib A.,** Simon I (2023) Characterization of Unidirectional Replication Forks in the Mouse Genome. *Int J Mol Sci* **24**:9611
57. **Bou-Nader C.,** Bothra A., Garboczi D.N., Leppla S.H., Zhang J (2022) Structural basis of R-loop recognition by the S9.6 monoclonal antibody. *Nat Commun* **13**:1641
58. **Hyrien O** (2015) Peaks cloaked in the mist: the landscape of mammalian replication origins. *J Cell Biol* **208**:147-160
59. **Tian M.,** et al. (2024) Integrative analysis of DNA replication origins and ORC-/MCM-binding sites in human cells reveals a lack of overlap. *eLife* **12**:RP89548 https://doi.org/10.7554/eLife.89548
60. **Marques C.A.,** Dickens N.J., Paape D., Campbell S.J., McCulloch R (2015) Genome-wide mapping reveals single-origin chromosome replication in *Leishmania*, a eukaryotic microbe. *Genome Biol* **16**
61. **Marques C.A.,** McCulloch R (2018) Conservation and Variation in Strategies for DNA Replication of Kinetoplastid Nuclear Genomes. *Curr Genomics* **19**:98-109
62. **Totañes F.I.G.,** et al. (2023) A genome-wide map of DNA replication at single-molecule resolution in the malaria parasite Plasmodium falciparum. *Nucleic Acids Res* **51**:2709-2724
63. **Segal E.,** Widom J (2009) Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr Opin Struct Biol* **19**:65-71
64. **Xu J.,** et al. (2012) Genome-wide identification and characterization of replication origins by deep sequencing. *Genome Biol* **13**
65. **Wirtz E.,** Leal S., Ochatt C., 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
66. **Hirumi H.,** 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
67. **Cayrou C.,** Grégoire D., Coulombe P., Danis E., Méchali M (2012) Genome-scale identification of active DNA replication origins. *Methods* **57**:158-164
68. **Martin M** (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**:10-12
69. **Langmead B.,** Wilks C., Antonescu V., Charles R (2019) Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **35**:421-432
70. **Danecek P.,** et al. (2021) Twelve years of SAMtools and BCFtools. *Gigascience* **10**:giab008
71. **Quinlan A.R.,** Hall I.M (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841-842
72. **Robinson J.T.,** et al. (2011) Integrative genomics viewer. *Nat Biotechnol* **29**:24-26

73. **Ramírez F.,** et al. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**:W160-165

74. **Bailey T.L.,** 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

75. **Heinz S.,** et al. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**:576-589

76. **Amos B.,** et al. (2022) VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Res* **50**:D898-D911

Peer reviews

Reviewer #1 (Public review):

In this paper, Stanojcic and colleagues attempt to map sites of DNA replication initiation in the genome of the African trypanosome, Trypanosoma brucei. Their approach to this mapping is to isolate 'short-nascent strands' (SNSs), a strategy adopted previously in other eukaryotes (including in the related parasite Leishmania major), which involves isolation of DNA molecules whose termini contain replication-priming RNA. By mapping the isolated and sequenced SNSs to the genome (SNS-seq), the authors suggest that they have identified origins, which they localise to intergenic (strictly, inter-CDS) regions within polycistronic transcription units and suggest display very extensive overlap with previously mapped R-loops in the same loci. Finally, having defined locations of SNS-seq mapping, they suggest they have identified G4 and nucleosome features of origins, again using previously generated data. Though there is merit in applying a new approach to understand DNA replication initiation in T. brucei, where previous work has used MFA-seq and ChIP of a subunit of the Origin Replication Complex (ORC), there are two significant deficiencies in the study that must be addressed to ensure rigour and accuracy.

(1) The suggestion that the SNS-seq data is mapping DNA replication origins that are present in inter-CDS regions of the polycistronic transcription units of T. brucei is novel and does not agree with existing data on the localisation of ORC1/CDC6, and it is very unclear if it agrees with previous mapping of DNA replication by MFA-seq due to the way the authors have presented this correlation. For these reasons, the findings essentially rely on a single experimental approach, which must be further tested to ensure SNS-seq is truly detecting origins. Indeed, in this regard, the very extensive overlap of SNS-seq signal with RNA-DNA hybrids should be tested further to rule out the possibility that the approach is mapping these structures and not origins.

(2) The authors' presentation of their SNS-seq data is too limited and therefore potentially provides a misleading view of DNA replication in the genome of T. brucei. The work is presented through a narrow focus on SNS-seq signal in the inter-CDS regions within polycistronic transcription units, which constitute only part of the genome, ignoring both the transcription start and stop sites at the ends of the units and the large subtelomeres, which are mainly transcriptionally silent. The authors must present a fuller and more balanced view of SNS-seq mapping across the whole genome to ensure full understanding and clarity.

https://doi.org/10.7554/eLife.108143.1.sa1

Reviewer #2 (Public review):

Summary:

Stanojcic et al. investigate the origins of DNA replication in the unicellular parasite Trypanosoma brucei. They perform two experiments, stranded SNS-seq and DNA molecular

combing. Further, they integrate various publicly available datasets, such as G4-seq and DRIP-seq, into their extensive analysis. Using this data, they elucidate the structure of the origins of replication. In particular, they find various properties located at or around origins, such as polynucleotide stretches, G-quadruplex structures, regions of low and high nucleosome occupancy, R-loops, and that origins are mostly present in intergenic regions. Combining their population-level SNS-seq and their single-molecule DNA molecular combing data, they elucidate the total number of origins as well as the number of origins active in a single cell.

Strengths:

(1) A very strong part of this manuscript is that the authors integrate several other datasets and investigate a large number of properties around origins of replication. Data analysis clearly shows the enrichment of various properties at the origins, and the manuscript concludes with a very well-presented model that clearly explains the authors' understanding and interpretation of the data.

(2) The DNA combing experiment is an excellent orthogonal approach to the SNS-seq data. The authors used the different properties of the two experiments (one giving location information, one giving single-molecule information) well to extract information and contrast the experiments.

(3) The discussion is exemplary, as the authors openly discuss the strengths and weaknesses of the approaches used. Further, the discussion serves its purpose of putting the results in both an evolutionary and a trypanosome-focused context.

Weaknesses:

I have major concerns about the origin of replication sites determined from the SNS-seq data. As a caveat, I want to state that, before reading this manuscript, SNS-seq was unknown to me; hence, some of my concerns might be misplaced.

(1) I do not understand why SNS-seq would create peaks. Replication should originate in one locus, then move outward in both directions until the replication fork moving outward from another origin is encountered. Hence, in an asynchronous population average measurement, I would expect SNS data to be broad regions of + and -, which, taken together, cover the whole genome. Why are there so many regions not covered at all by reads, and why are there such narrow peaks?

(2) I am concerned that up to 96% percent of all peaks are filtered away. If there is so much noise in the data, how can one be sure that the peaks that remain are real? Specifically, if the authors placed the same number of peaks as was measured randomly in intergenic regions, would 4% of these peaks pass the filtering process by chance?

(3) There are 3 previous studies that map origins of replication in T. brucei. Devlin et al. 2016, Tiengwe et al. 2012, and Krasiļņikova et al. 2025 (*https://doi.org/10.1038/s41467-025-56087-3*), all with a different technique: MFA-seq. All three previous studies mostly agree on the locations and number of origins. The authors compared their results to the first two, but not the last study; they found that their results are vastly different from the previous studies (see Supplementary Figure 8A). In their discussion, the authors defend this discrepancy mostly by stating that the discrepancy between these methods has been observed in other organisms. I believe that, given the situation that the other studies precede this manuscript, it is the authors' duty to investigate the differences more than by merely pointing to other organisms. A conclusion should be reached on why the results are different, e.g., by orthogonally validating origins absent in the previous studies.

(4) Some patterns that were identified to be associated with origins of replication, such as G-quadruplexes and nucleosomes phasing, are known to be biases of SNS-seq (see Foulk et al.

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. 2015;25(5):725-735. doi:10.1101/gr.183848.114).

Are the claims well substantiated?:

My opinion on whether the authors' results support their conclusions depends on whether my concerns about the sites determined from the SNS-seq data can be dismissed. In the case that these concerns can be dismissed, I do think that the claims are compelling.

Impact:

If the origins of replication prove to be distributed as claimed, this study has the potential to be important for two fields. Firstly, in research focused on T. brucei as a disease agent, where essential processes that function differently than in mammals are excellent drug targets. Secondly, this study would impact basic research analyzing DNA replication over the evolutionary tree, where T. brucei can be used as an early-divergent eukaryotic model organism.

https://doi.org/10.7554/eLife.108143.1.sa0

Author response:

> ***eLife Assessment***
>
> *The authors use sequencing of nascent DNA (DNA linked to an RNA primer, "SNS-Seq") to localise DNA replication origins in Trypanosoma brucei, so this work will be of interest to those studying either Kinetoplastids or DNA replication. The paper presents the SNS-seq results for only part of the genome, and there are significant discrepancies between the SNS-Seq results and those from other, previously-published results obtained using other origin mapping methods. The reasons for the differences are unknown and from the data available, it is not possible to assess which origin-mapping method is most suitable for origin mapping in T. brucei. Thus at present, the evidence that origins are distributed as the authors claim - and not where previously mapped - is inadequate.*

We would like to clarify a few points regarding our study. Our primary objective was to characterise the topology and genome-wide distribution of short nascent-strand (SNS) enrichments. The stranded SNS-seq approach provides the high strand-specific resolution required to analyse origins. The observation that SNS-seq peaks (potential origins) are most frequently found in intergenic regions is not an artefact of analysing only part of the genome; rather, it is a result of analysing the entire genome.

We agree that orthogonal validation is necessary. However, neither MFA-seq nor TbORC1/CDC6 ChIP-on-chip has yet been experimentally validated as definitive markers of origin activity in T. brucei, nor do they validate each other.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public review):***
>
> *In this paper, Stanojcic and colleagues attempt to map sites of DNA replication initiation in the genome of the African trypanosome, Trypanosoma brucei. Their approach to this mapping is to isolate 'short-nascent strands' (SNSs), a strategy adopted previously in other eukaryotes (including in the related parasite Leishmania major), which involves isolation of DNA molecules whose termini contain replication-priming RNA. By mapping the isolated and sequenced SNSs to the genome (SNS-seq), the authors suggest that they have identified origins, which they localise to intergenic (strictly, inter-CDS) regions within*

> *polycistronic transcription units and suggest display very extensive overlap with previously mapped R-loops in the same loci. Finally, having defined locations of SNS-seq mapping, they suggest they have identified G4 and nucleosome features of origins, again using previously generated data.*
>
> *Though there is merit in applying a new approach to understand DNA replication initiation in T. brucei, where previous work has used MFA-seq and ChIP of a subunit of the Origin Replication Complex (ORC), there are two significant deficiencies in the study that must be addressed to ensure rigour and accuracy.*
>
> *(1) The suggestion that the SNS-seq data is mapping DNA replication origins that are present in inter-CDS regions of the polycistronic transcription units of T. brucei is novel and does not agree with existing data on the localisation of ORC1/CDC6, and it is very unclear if it agrees with previous mapping of DNA replication by MFA-seq due to the way the authors have presented this correlation. For these reasons, the findings essentially rely on a single experimental approach, which must be further tested to ensure SNS-seq is truly detecting origins. Indeed, in this regard, the very extensive overlap of SNS-seq signal with RNA-DNA hybrids should be tested further to rule out the possibility that the approach is mapping these structures and not origins.*
>
> *(2) The authors' presentation of their SNS-seq data is too limited and therefore potentially provides a misleading view of DNA replication in the genome of T. brucei. The work is presented through a narrow focus on SNS-seq signal in the inter-CDS regions within polycistronic transcription units, which constitute only part of the genome, ignoring both the transcription start and stop sites at the ends of the units and the large subtelomeres, which are mainly transcriptionally silent. The authors must present a fuller and more balanced view of SNS-seq mapping across the whole genome to ensure full understanding and clarity.*

Regarding comparisons with previous work:

Two other attempts to identify origins in *T. brucei* —ORC1/CDC6 binding sites (ChIP-on-chip, PMID: 22840408) and MFA-seq (PMID: 22840408, 27228154)—were both produced by the McCulloch group. These methods do not validate each other; in fact, MFA-seq origins overlap with only 4.4% of the 953 ORC1/CDC6 sites (PMID: 29491738). Therefore, low overlap between SNS-seq peaks and ORC1/CDC6 sites cannot disqualify our findings. Similar low overlaps are observed in other parasites (PMID: 38441981, PMID: 38038269, PMID: 36808528) and in human cells (PMID: 38567819).

We also would like to emphasize that the ORC1/CDC6 dataset originally published (PMID: 22840408) is no longer available; only a re-analysis by TritrypDB exists, which differs significantly from the published version (personal communication from Richard McCulloch). While the McCulloch group reported a predominant localization of ORC1/CDC6 sites within SSRs at transcription start and termination regions, our re-analysis indicates that only 10.3% of TbORC1/CDC6-12Myc sites overlapped with 41.8% of SSRs.

MFA-seq does not map individual origins, it rather detects replicated genomic regions by comparing DNA copy number between S- and G1-phases of the cell cycle (PMID: 36640769; PMID: 37469113; PMID: 36455525). The broad replicated regions (0.1–0.5 Mbp) identified by MFA-seq in *T. brucei* are likely to contain multiple origins, rather than just one. In that sense we disagree with the McCulloch's group who claimed that there is a single origin per broad peak. Our analysis shows that up to 50% of the origins detected by stranded SNS-seq locate within broad MFA-seq regions. The methodology used by McCulloch's group to infer single origins from MFA-seq regions has not been published or made available, as well as the precise position of these regions, making direct comparison difficult.

Finally, the genomic features we describe—poly(dA/dT) stretches, G4 structures and nucleosome occupancy patterns—are consistent with origin topology described in other organisms.

On the concern that SNS-seq may map RNA-DNA hybrids rather than replication origins: Isolation and sequencing of short nascent strands (SNS) is a well-established and widely used technique for high-resolution origin mapping. This technique has been employed for decades in various laboratories, with numerous publications documenting its use. We followed the published protocol for SNS isolation (Cayrou et al., Methods, 2012, PMID: 22796403). RNA-DNA hybrids cannot persist through the multiple denaturation steps in our workflow, as they melt at 95°C (Roberts and Crothers, *Science*, 1992; PMID: 1279808). Even in the unlikely event that some hybrids remained, they would not be incorporated into libraries prepared using a single-stranded DNA protocol and therefore would not be sequenced (see Figure 1B and Methods).

Furthermore, our analysis shows that only a small proportion (1.7%) of previously reported RNA-DNA hybrids overlap with SNS-seq origins. It is important to note that RNA-primed nascent strands naturally form RNA-DNA hybrids during replication initiation, meaning the enrichment of RNA-DNA hybrids near origins is both expected and biologically relevant.

On the claim that our analysis focuses narrowly on inter-CDS regions and ignores other genomic compartments: this is incorrect. We mapped and analyzed stranded SNS-seq data across the entire genome of *T. brucei* 427 wild-type strain (Müller et al., *Nature*, 2018; PMID: 30333624), including both core and subtelomeric regions. Our findings indicate that most origins are located in intergenic regions, but all analyses were performed using the full set of detected origins, regardless of location.

We did not ignore transcription start and stop sites (TSS/TTS). The manuscript already includes origin distribution across genomic compartments as defined by TriTrypDB (Fig. 2C) and addresses overlap with TSS, TTS and HT in the section “Spatial coordination between the activity of the origin and transcription”. While this overlap is minimal, we have included metaplots in the revised manuscript for clarity.

> ***Reviewer #2 (Public review):***
>
> *Summary:*
>
> *Stanojcic et al. investigate the origins of DNA replication in the unicellular parasite Trypanosoma brucei. They perform two experiments, stranded SNS-seq and DNA molecular combing. Further, they integrate various publicly available datasets, such as G4-seq and DRIP-seq, into their extensive analysis. Using this data, they elucidate the structure of the origins of replication. In particular, they find various properties located at or around origins, such as polynucleotide stretches, G-quadruplex structures, regions of low and high nucleosome occupancy, R-loops, and that origins are mostly present in intergenic regions. Combining their population-level SNS-seq and their single-molecule DNA molecular combing data, they elucidate the total number of origins as well as the number of origins active in a single cell.*
>
> *Strengths:*
>
> *(1) A very strong part of this manuscript is that the authors integrate several other datasets and investigate a large number of properties around origins of replication. Data analysis clearly shows the enrichment of various properties at the origins, and the manuscript concludes with a very well-presented model that clearly explains the authors' understanding and interpretation of the data.*

We sincerely thank you for this positive feedback.

> *(2) The DNA combing experiment is an excellent orthogonal approach to the SNS-seq data. The authors used the different properties of the two experiments (one giving location information, one giving single-molecule information) well to extract information and contrast the experiments.*

Thank you very much for this remark.

> *(3) The discussion is exemplary, as the authors openly discuss the strengths and weaknesses of the approaches used. Further, the discussion serves its purpose of putting the results in both an evolutionary and a trypanosome-focused context.*

Thank you for appreciating our discussion.

> *Weaknesses:*
>
> *I have major concerns about the origin of replication sites determined from the SNS-seq data. As a caveat, I want to state that, before reading this manuscript, SNS-seq was unknown to me; hence, some of my concerns might be misplaced.*
>
> *(1) I do not understand why SNS-seq would create peaks. Replication should originate in one locus, then move outward in both directions until the replication fork moving outward from another origin is encountered. Hence, in an asynchronous population average measurement, I would expect SNS data to be broad regions of + and -, which, taken together, cover the whole genome. Why are there so many regions not covered at all by reads, and why are there such narrow peaks?*

Thank you for asking these questions. As you correctly point out, replication forks progress in both directions from their origins and ultimately converge at termination sites. However, the SNS-seq method specifically isolates short nascent strands (SNSs) of 0.5–2.5 kb using a sucrose gradient. These short fragments are generated immediately after origin firing and mark the sites of replication initiation, rather than the entire replicated regions. Consequently: (i) SNS-seq does not capture long replication forks or termination regions, only the immediate vicinity of origins. (ii) The narrow peaks indicate the size of selected SNSs (0.5–2.5 kb) and the fact that many cells initiate replication at the same genomic sites, leading to localized enrichment. (iii) Regions without coverage refer to genomic areas that do not serve as efficient origins in the analyzed cell population. Thus, SNS-seq is designed to map origin positions, but not the entire replicated regions.

> *(2) I am concerned that up to 96% percent of all peaks are filtered away. If there is so much noise in the data, how can one be sure that the peaks that remain are real? Specifically, if the authors placed the same number of peaks as was measured randomly in intergenic regions, would 4% of these peaks pass the filtering process by chance?*

Maintaining the strandness of the sequenced DNA fibres enabled us to filter the peaks, thereby increasing the probability that the filtered peak pairs corresponded to origins. 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.

As suggested by the reviewer, we tested whether randomly placed plus and minus peaks could reproduce the number of filter-passing peaks using the same bioinformatics workflow. Only 1–6% of random peaks passed the filters, compared with 4–12% in our experimental data, resulting in about 50% fewer selected regions (origins). Moreover, the “origins” from random peaks showed 0% reproducibility across replicates, whereas the experimental data showed 7–64% reproducibility. These results indicate that the retainee peaks are highly

unlikely to arise by chance and support the specificity of our approach. Thank you for this suggestion.

> *(3) There are 3 previous studies that map origins of replication in T. brucei. Devlin et al. 2016, Tiengwe et al. 2012, and Krasiļnikova et al. 2025 (https://doi.org/10.1038/s41467-025-56087-3), all with a different technique: MFA-seq. All three previous studies mostly agree on the locations and number of origins. The authors compared their results to the first two, but not the last study; they found that their results are vastly different from the previous studies (see Supplementary Figure 8A). In their discussion, the authors defend this discrepancy mostly by stating that the discrepancy between these methods has been observed in other organisms. I believe that, given the situation that the other studies precede this manuscript, it is the authors' duty to investigate the differences more than by merely pointing to other organisms. A conclusion should be reached on why the results are different, e.g., by orthogonally validating origins absent in the previous studies.*

The MFA-seq data for *T. brucei* were published in two studies by McCulloch's group: Tiengwe et al. (2012) using TREU927 PCF cells, and Devlin et al. (2016) using PCF and BSF Lister427 cells. In Krasilnikova et al. (2025), previously published MFA-seq data from Devlin et al. were remapped to a new genome assembly without generating new MFA-seq data, which explains why we did not include that comparison.

Clarifying the differences between MFA-seq and our stranded SNS-seq data is essential. MFA-seq and SNS-seq interrogate different aspects of replication. SNS-seq is a widely used, high-resolution method for mapping individual replication origins, whereas MFA-seq detects replicated regions by comparing DNA copy number between S and G1 phases. MFA-seq identified broad replicated regions (0.1–0.5 Mb) that were interpreted by McCulloch's group as containing a single origin. We disagree with this interpretation and consider that there are multiple origins in each broad peaks; theoretical considerations of replication timing indicate that far more origins are required for complete genome duplication during the short S-phase. Once this assumption is reconsidered, MFA-seq and SNS-seq results become complementary: MFA-seq identifies replicated regions, while SNS-seq pinpoints individual origins within those regions. Our analysis revealed that up to 50% of the origins detected by stranded SNS-seq were located within the broad MFA peaks. This pattern—broad MFA-seq regions containing multiple initiation sites—has also recently been found in *Leishmania* by McCulloch's team using nanopore sequencing (PMID: 26481451). Nanopore sequencing showed numerous initiation sites within MFA-seq regions and additional numerous sites outside these regions in asynchronous cells, consistent with what we observed using stranded SNS-seq in *T. brucei*. We will expand our discussion and conclude that the discrepancy arises from methodological differences and interpretation. The two approaches provide complementary insights into replication dynamics, rather than 'vastly different' results.

We recognize the importance of validating our results in future using an alternative mapping method and functional assays. However, it is important to emphasize that stranded SNS-seq is an origin mapping technique with a very high level of resolution. This technique can detect regions between two divergent SNS peaks, which should represent regions of DNA replication initiation. At present, no alternative technique has been developed that can match this level of resolution.

> *(4) Some patterns that were identified to be associated with origins of replication, such as G-quadruplexes and nucleosomes phasing, are known to be biases of SNS-seq (see Foulk et al. 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. 2015;25(5):725-735. doi:10.1101/gr.183848.114).*

It is important to note that the conditions used in our study differ significantly from those applied in the Foulk et al. Genome Res. 2015. We used SNS isolation and enzymatic treatments as described in previous reports (Cayrou, C. et al. Genome Res, 2015 and Cayrou, C et al. Methods, 2012). Here, we enriched the SNS by size on a sucrose gradient and then treated this SNS-enriched fraction with high amounts of repeated λ-exonuclease treatments (100u for 16h at 37oC - see Methods). In contrast, Foulk et al. used sonicated total genomic DNA for origin mapping, without enrichment of SNS on a sucrose gradient as we did, and then they performed a λ-exonuclease treatment. A previous study (Cayrou, C. et al. Genome Res, 2015, Figure S2, which can be found at *https://genome.cshlp.org/content/25/12/1873/suppl/DC1*) has shown that complete digestion of G4-rich DNA sequences is achieved under the conditions we used.

Furthermore, the SNS depleted control (without RNA) was included in our experimental approach. This control represents all molecules that are difficult to digest with lambda exonuclease, including G4 structures. Peak calling was performed against this background control, with the aim of removing false positive peaks resulting from undigested DNA structures. We explained better this step in the revised manuscript.

The key benefit of our study is that the orientation of the enrichments (peaks) remains consistent throughout the sequencing process. We identified an enrichment of two divergent strands synthesised on complementary strands containing G4s. These two divergent strands themselves do not, however, contain G4s (see Fig. 8 for the model). Therefore, the enriched molecules detected in our study do not contain G4s. They are complementary to the strands enriched with G4s. This means that the observed enrichment of

G4s cannot be an artefact of the enzymatic treatments used in this study. We added this part in the discussion of the revised manuscript.

We also performed an additional control which is not mentioned in the manuscript. In parallel with replicating cells, we isolated the DNA from the stationary phase of growth, which primarily contains non-replicating cells. Following the three λ-exonuclease treatments, there was insufficient DNA remaining from the stationary phase cells to prepare the libraries for sequencing. This control strongly indicated that there was little to no contaminating DNA present with the SNS molecules after λ-exonuclease enrichment.

https://doi.org/10.7554/eLife.108143.1.sa3