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

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

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

[2] Trypanosoma unit, Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium

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

[#] Authors contributed to the manuscript equally

* Corresponding authors: slavica.stanojcic@umontpellier.fr, yvon.sterkers@umontpellier.fr

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 RNA: DNA hybrids. 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.

Key words

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

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 [41–43]. 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

70 preserves their orientation. This stranded SNS-seq method allowed the mapping of origins
71 between two divergently oriented SNS peaks and the discovery of potential DNA replication
72 start sites. We then conducted an extensive bioinformatics analysis of the mapped origins
73 and compared our findings with previously published scientific datasets. This enabled us to
74 identify several key components of origins and to propose a model of an origin in *T. brucei*.

75 **RESULTS**

76 **Origin mapping by stranded SNS-seq.**

77 Stranded SNS-seq method was employed to map the origins in two distinct cell types of *T.*
78 *brucei*, representing two life cycle stages of this parasite: PCF and BSF cells. The experimental
79 workflow for purifying and enriching SNS, together with the corresponding control, is
80 illustrated in Figure 1A. SNS-enriched samples were prepared through a two-stage process.
81 SNS molecules were enriched first, by size-selection (0.5-2.5 kb), and second, by removing
82 broken DNA molecules by λ-exonuclease treatments. SNS-depleted control, prepared in
83 parallel with every SNS-enriched sample, underwent an additional RNase treatment before λ-
84 exonuclease treatment to remove RNA, including RNA primers from the 5' end of the SNS
85 (Methods and Figure 1A). SNS-depleted control contained molecules resistant to RNase
86 and/or λ-exonuclease treatment, representing all molecules that are difficult to digest with
87 lambda exonuclease. This control was utilised as a background control. Then, stranded
88 libraries were prepared for high-throughput sequencing (Figure 1B and Methods). Paired-
89 end sequencing of an SNS-enriched samples and an SNS-depleted controls was performed
90 on three biological replicates of PCF and BSF cells. Read counts after sequencing and each
91 analysis step are shown for each sample in Supplementary Table 1 and Supplementary Figure
92 1A.

93 The bioinformatics workflow used for origin selection is schematically presented in Figure 1C
94 and Methods. In brief, after the alignment and strand-specific sorting of the reads, two steps
95 of peak selection were performed. The first step was strand-specific peak calling against
96 SNS-depleted control as a background control. The aim of this step was to eliminate false
97 positive peaks resulting from undigested DNA structures, such as G4s and other structures
98 that could withstand λ-exonuclease treatment. The second step, called "peak filtering",
99 identified peaks pairs consistent with the divergent orientation of SNS molecules at active
100 origins (Figure 1C). Therefore, only peak pairs with a minus peak in front of a plus peak, and
101 both in proximity (Methods and Figure 1C and 1D) were retained as positives, while 88-96%
102 of the called peaks were removed in the "peak filtering" step (Supplementary Figure 1B and
103 1C). Here, we used information about the divergent orientation of SNS, a key feature of
104 active origins, to increase the specificity of origin mapping. Ultimately, the origins were
105 identified as genomic regions situated between the inner borders of filtered peak pairs
106 (Figure 1D).

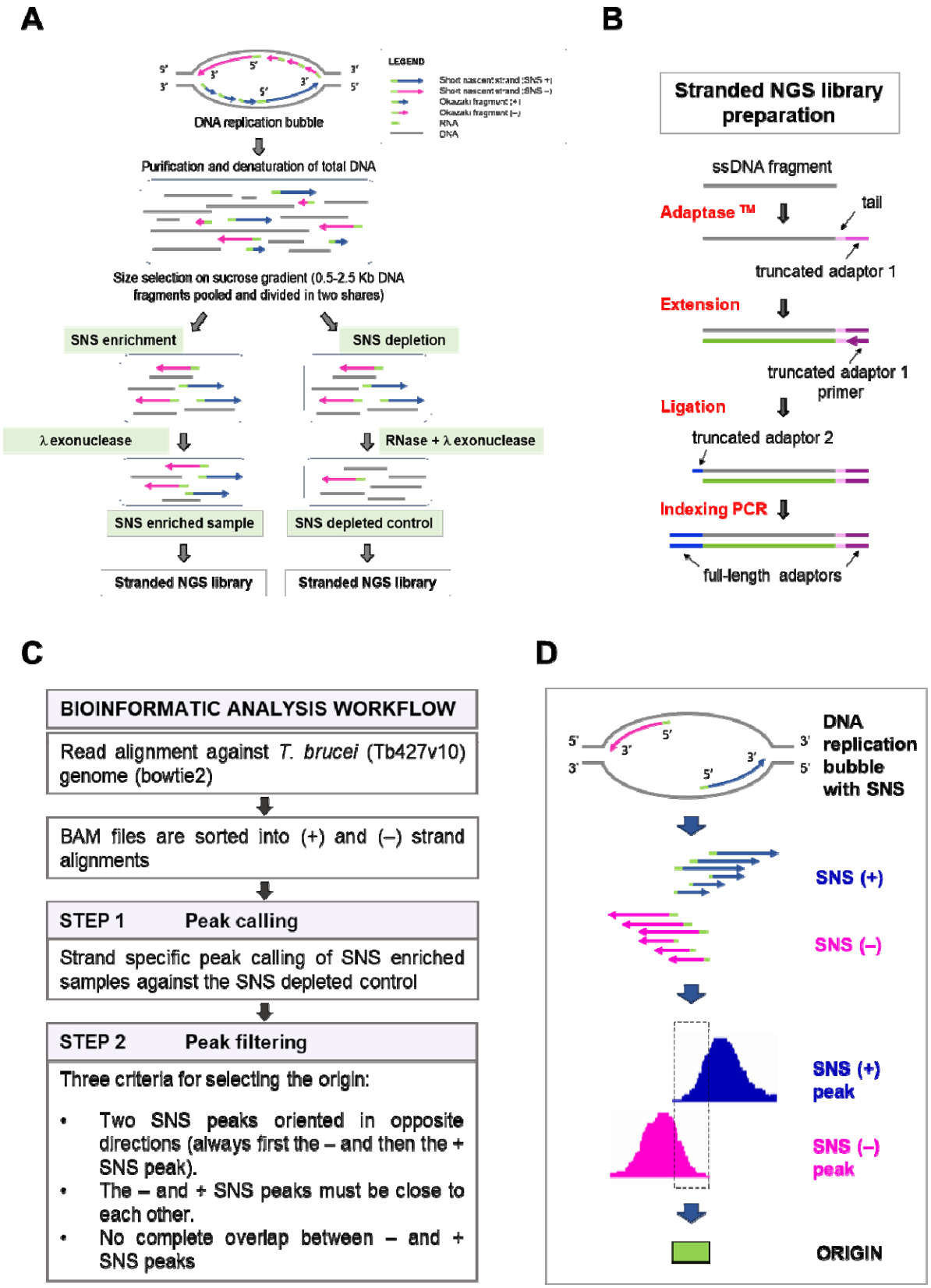


Figure 1. Experimental protocol and bioinformatics workflow for selecting the origins.

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 stranded SNS-seq data were mapped across the entire genome, including 11 core regions and all non-core regions (subtelomeric and 'unitig') [44]. 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. After mapping, we merged the origins from the three replicates, identifying a total of 914 origins in BSF cells and 549 in PCF cells. Of the origins in BSF cells, 742 (81.2%) were mapped to the core genomic region and 172 (18.8%) to the non-core regions. In PCF cells, 492 origins (89.6%) were mapped to the core genomic region and 56 (10.4%) to the non-core regions (Figure 2B). Therefore, although these regions present more than half of the genomic sequence (Figure 2B), we mapped fewer origins in non-core regions.

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 2C). 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 2C). 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 2D). 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 2D).

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).

153 Empirical cumulative distribution function (ECDF) was calculated for both cell types, and a
154 statistically significant difference was confirmed between the origins and shuffled controls
155 clustering (Figure 2E). The results showed that 50% of the PCF origins were clustered within
156 11.5 kb regions (shuffled control median was 17 kb), while 50% of the BSF origins were
157 clustered within 6.5 kb regions (shuffled control median was 13 kb) (Figure 2E). This means
158 that origins tend to be activated within initiation zones or clusters rather than as solitary
159 initiation sites in *T. brucei*. This finding is consistent with previous research in other
160 organisms, which has shown that origins are unevenly distributed and tend to form clusters
161 of different lengths [20,22,25]. Next, the intersection of PCF and BSF origins revealed that 238
162 origins are in common (Supplementary Figure 2C). The number of origins in both cells was
163 compared; for this, the reads were normalised to have an equal number of mapped reads in
164 both cell types (Methods, Read sub-sampling). We showed that BSF cells contained almost
165 twice as many origins as PCF cells (BSF: 844 origins and PCF: 454 origins) (Supplementary
166 Figure 2D).

167 We then calculated the inter-origin distances (IODs) distribution at the population level for
168 both cell types. The calculated population-based IODs were shorter in BSF (median 15.3 kb)
169 compared to PCF cells (median 21.4 kb) and this difference was statistically significant
170 (Supplementary Figure 2E). Very long IODs were rarely observed, with maximum distances of
171 354 and 488 kilobases for PCF and BSF origins, respectively (Supplementary Figure 2E).

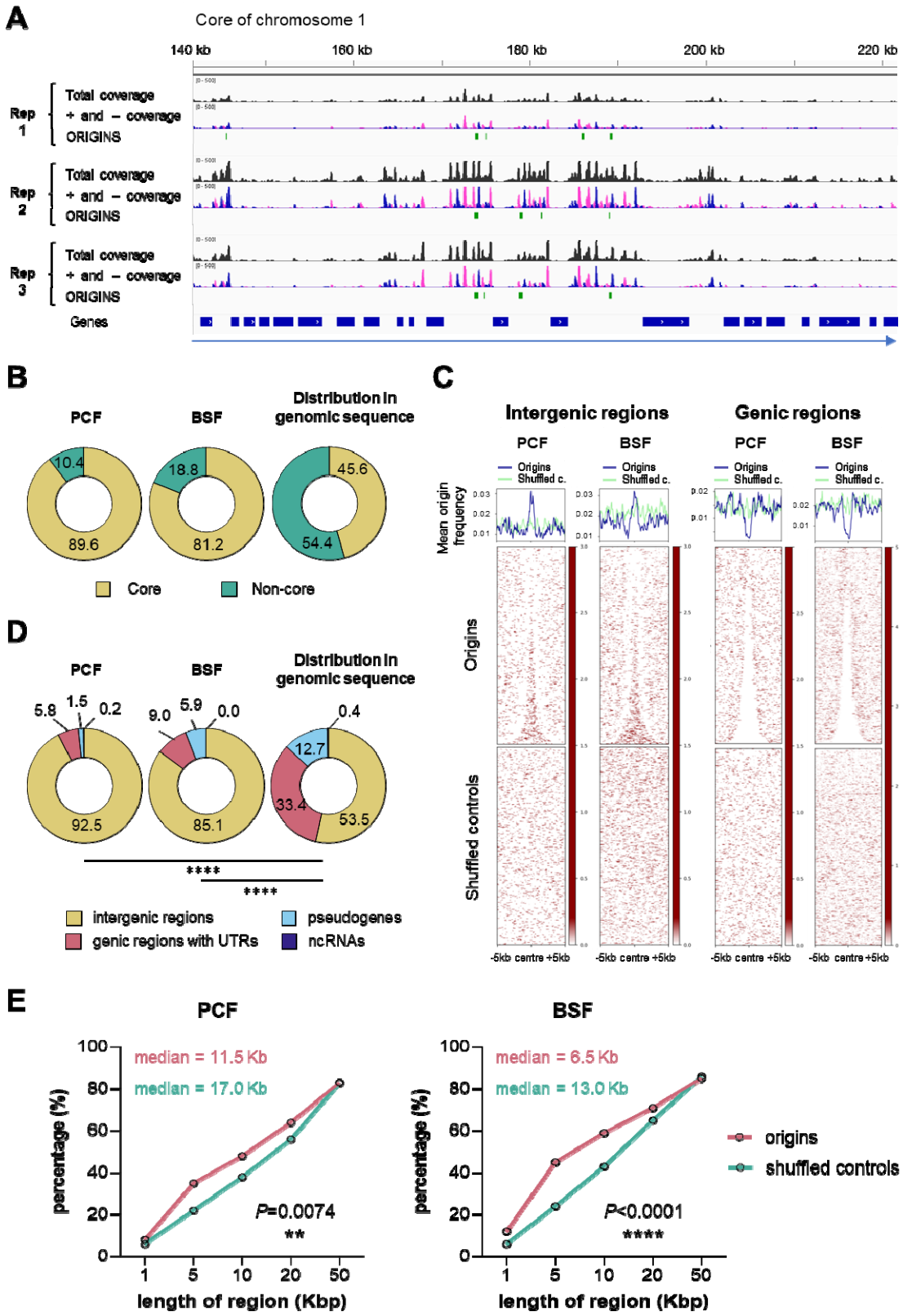


Figure 2. Genomic distribution of mapped origins.

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) The pie charts show the proportions of mapped origins within the core (yellow) and non-core (green) regions in PCF and BSF cells, as well as the distribution of these two regions within the genomic sequence. C) 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. D) The pie charts illustrate the proportions of the mapped origins of PCF and BSF cells within four indicated genomic regions, as well as the distribution of these four regions within the 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 [45]). 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 [46], 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

217 analysed DNA molecules but rather reflected genuine differences between the replication of
218 PCF and BSF cells.

219 The approximate number of active origins in a single cell can be estimated by dividing the
220 length of genomic sequences (~50 Mb) by the mean or median IODs obtained by DNA
221 combing. It was thus estimated that the range of active origins in a single cell was 279-329
222 for PCF and 228-235 for BSF cells (Supplementary Table 2). The stranded SNS-seq method
223 revealed 549 origins for PCF and 914 origins for BSF under stringent origin selection, which
224 represent a fraction of all origins. Despite the inability of either method to determine the
225 exact number of origins, the number of origins in a single cell tends to be smaller than in a
226 population of cells. This phenomenon, known as flexible origin usage, is a well-documented
227 feature of eukaryotic origins (reviewed in [45]).

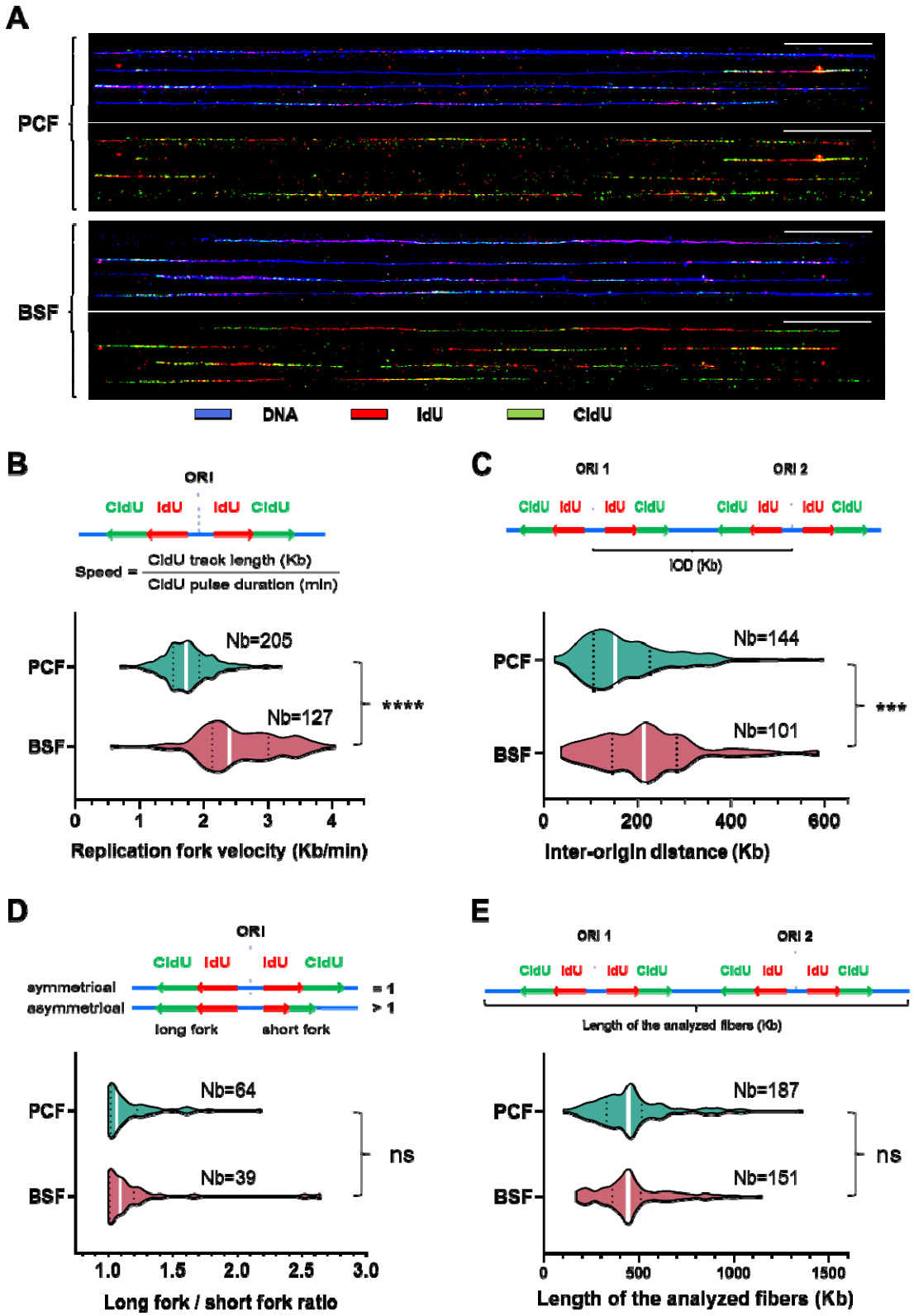


Figure 3. Comparison of DNA replication parameters between PCF and BSF cells by DNA molecular combing.

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.

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 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

276 consensus motif was identified (data not shown). The sequences surrounding the origins
277 were also analysed using the WebLogo software [47], and a graphical representation of
278 nucleotide occurrence unambiguously confirmed an unequal nucleotide distribution, with an
279 enrichment of A and G upstream and T and C downstream of the origins (Supplementary
280 Figure 3B). We then looked at the distribution of polynucleotides of different lengths around
281 the origins. Poly(dA)-rich sequences were enriched upstream and poly(dT)-rich sequences
282 downstream of the origins, a pattern absent in shuffled controls (Figure 4B, Supplementary
283 Figure 3C). Interestingly, the upstream region showed an enrichment of up to four
284 consecutive Gs, while the downstream region showed an enrichment of up to four
285 consecutive Cs (Figure 4B). We also checked nucleotide and polynucleotide distributions
286 around PCF and BSF origins separately and found that the distributions were similar in both
287 cell types and equivalent to the merged set of origins (Supplementary Figure 4A and
288 Supplementary Figure 4B). Accordingly, an origin can be defined as a genomic site with a
289 specific uneven distribution of nucleotides, characterised by an overrepresentation of
290 poly(dA)s and poly(dG)s upstream, and poly(dT)s and poly(dC)s downstream of the origin
291 centre.

292 Subsequently, the proportions of origins that lacked or possessed four As upstream and/or
293 four Ts downstream of the centre were calculated and compared with the intergenic shuffled
294 controls (Figure 4C, left graph). The same calculation was performed for eight As and Ts
295 (Figure 4C, right graph). The findings revealed that 1149 origins exhibited the presence of
296 four As upstream and four Ts downstream of the centre, representing a proportion of ~94%
297 compared to ~78% for the intergenic shuffled controls. Similarly, 450 origins (~37%)
298 possessed eight As upstream and eight Ts downstream of the centre, compared to only 114
299 (~9%) intergenic shuffled controls with this pattern. Accordingly, the distribution of poly(dA)
300 and poly(dT) observed for origins was found to be significantly different from that observed
301 for the intergenic shuffled controls (Figure 4C).

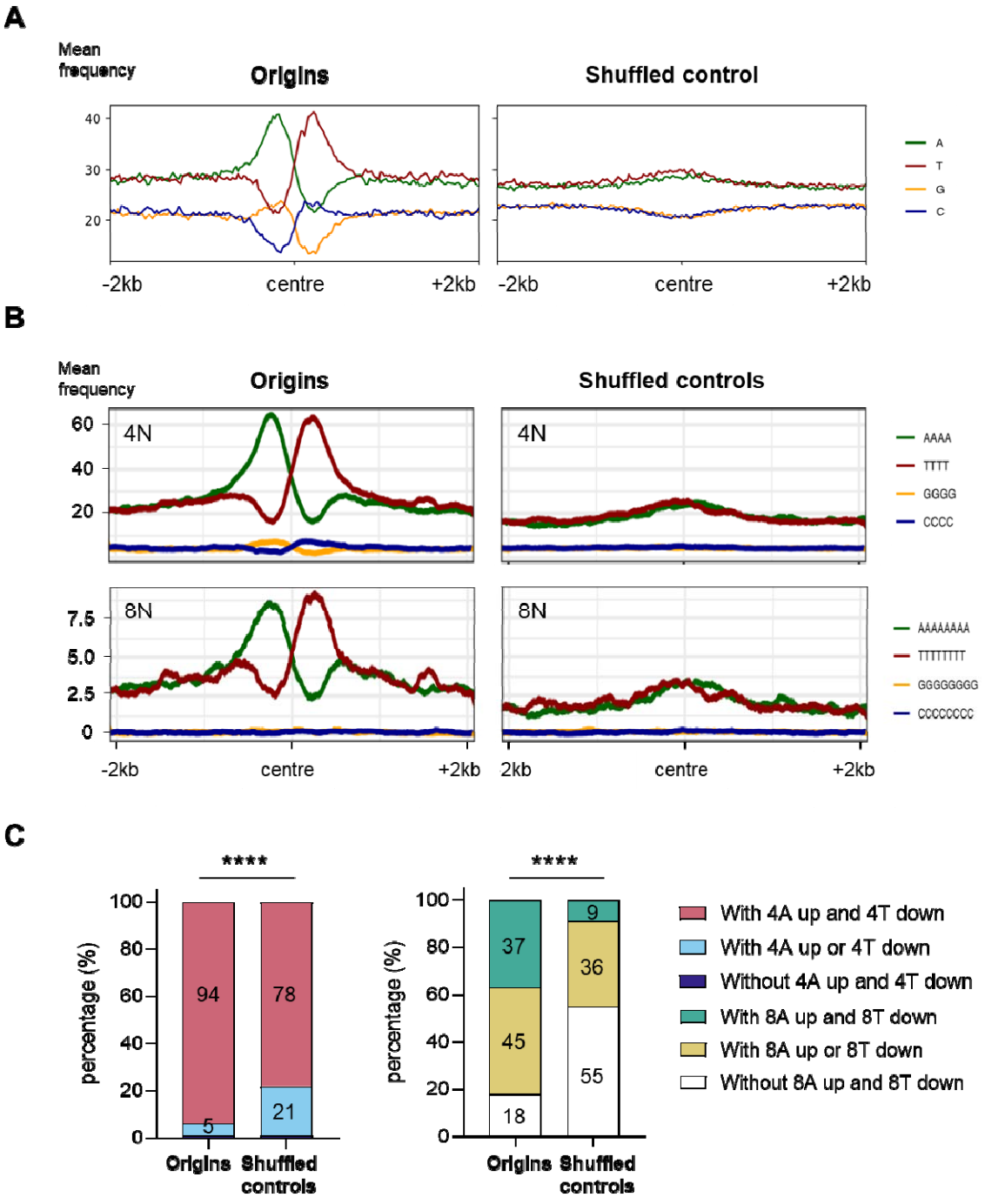


Figure 4. Spatial organisation of nucleotides and polynucleotides in the vicinity of origins.

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).

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 [48]. 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) [49]. 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 [49]. 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 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

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

353 Our results demonstrated that poly(dA) enriched sequences and G4s were enriched upstream
354 of origins on the plus strand, and downstream on the minus strand (Figure 4B and Figure
355 5A). Therefore, we wanted to determine the spatial relationship between these two elements.
356 To achieve this, the strand separated poly(dA) enriched sequences and the strand separated
357 experimental G4s [49] were intersected at centred origins and shuffled controls (Methods). The
358 results demonstrated that the peaks of the experimental G4s and poly(dA) sequences
359 overlapped near the centres of origins in a strand specific manner that was notably absent in
360 the intergenic shuffled controls (Figure 5D). Our findings revealed a previously unidentified
361 element of origin: the enrichment of poly(dA) and G4 structures exhibited an important
362 association upstream of the centre on the plus strand and downstream of origin on the
363 minus strand (Figure 5D). It was observed that the poly(dA) peaks exhibited summits
364 between ±100 to 300 nt, while the G4 peaks exhibited summits in the range of ±100 to 200
365 nt from the centres of origins (Figure 5D). Such regions were designated as poly(dA)/G4-
366 enriched elements of origins.

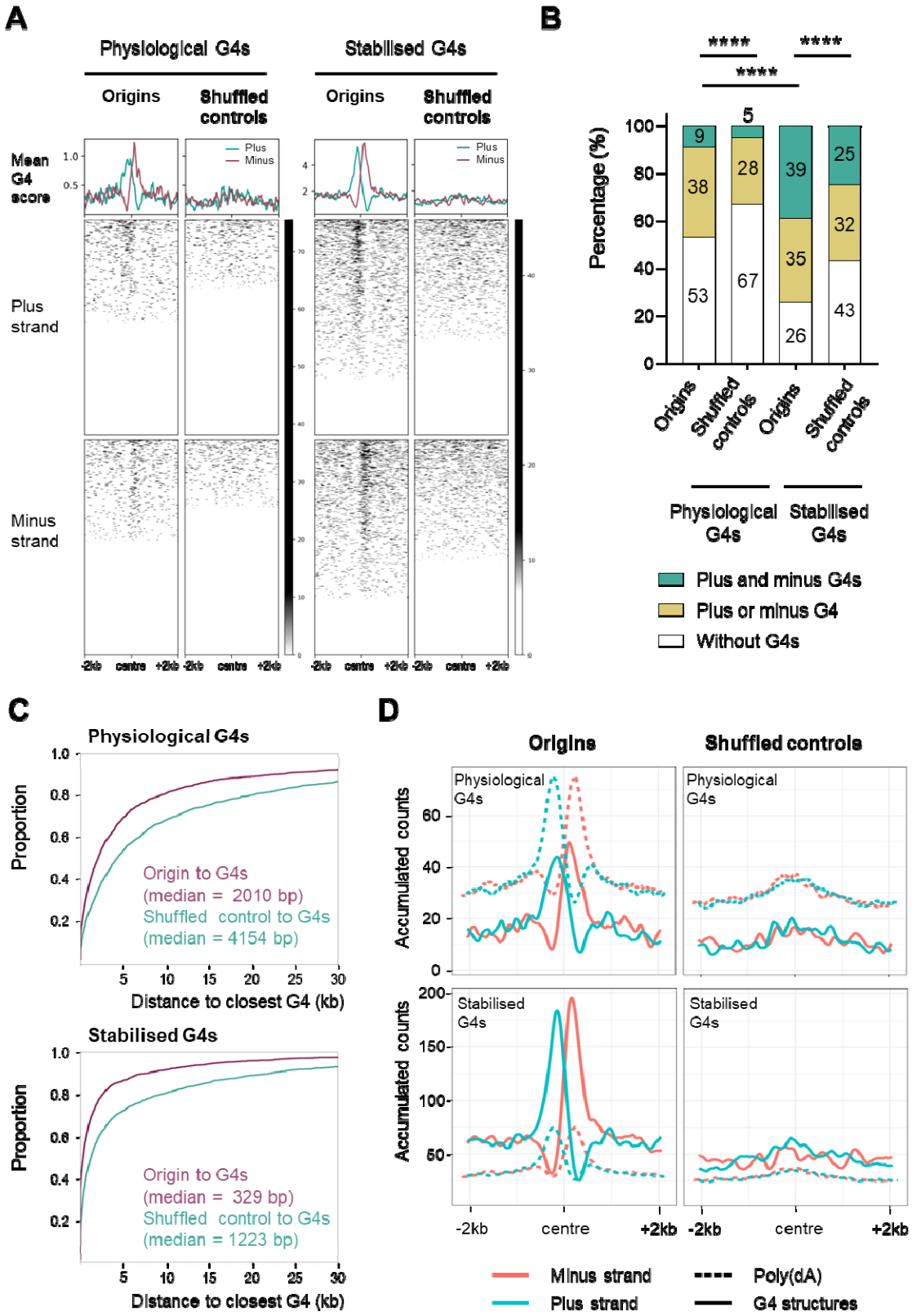


Figure 5. Distribution of G4 structures in the vicinity of origins.

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[49] were overlapped with origins mapped by stranded SNS-seq. Mean G4 score presents average G4 score [49] 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 [49] 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 [49]. 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) [49] (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).

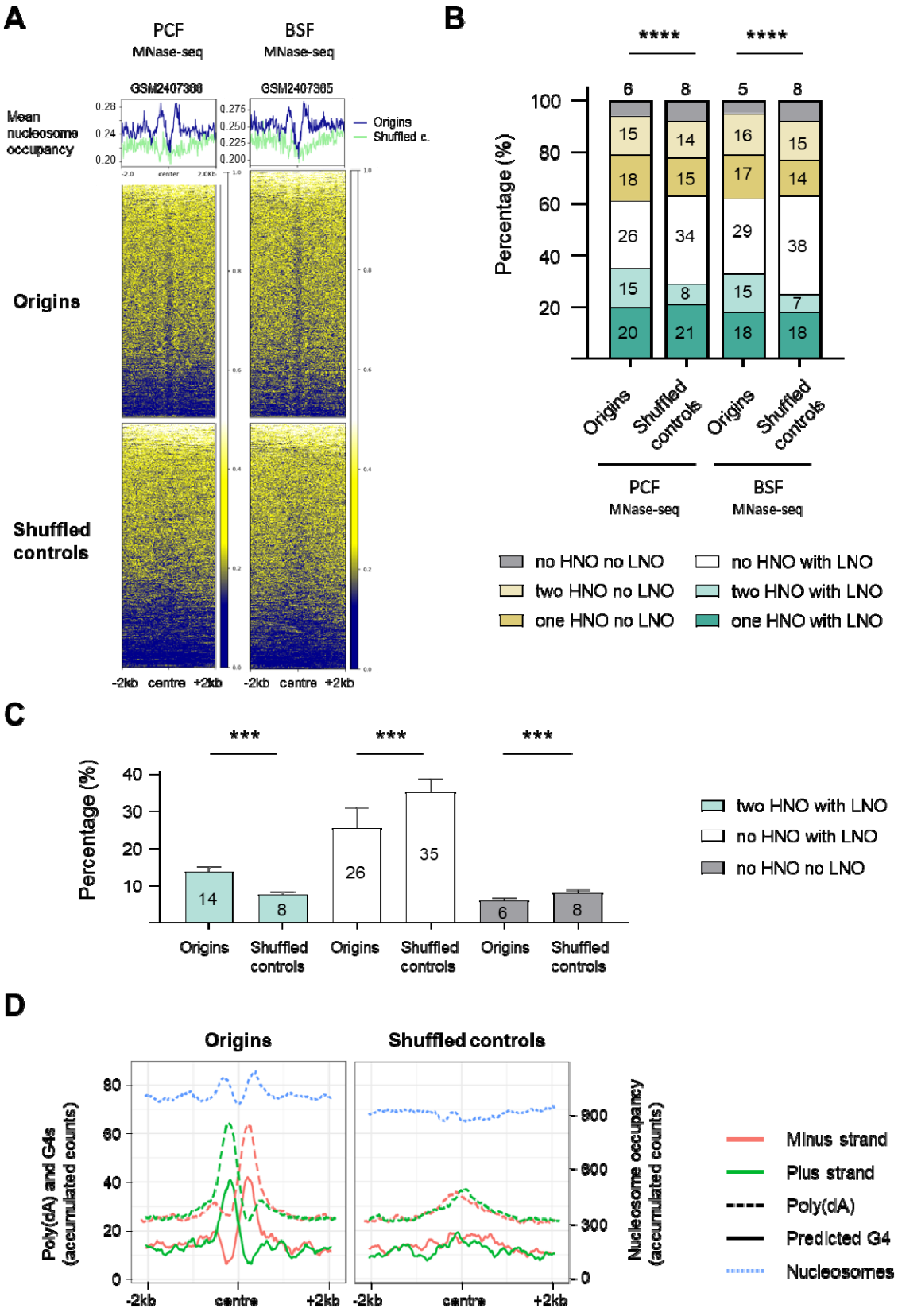
The origins exhibited varying nucleosome occupancy.

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* [50]. 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).

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 [50] (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

413 depleted in the origins compared to the shuffled controls (Figure 6C). Therefore, the active
414 origins had significantly more LNO flanked by two HNO, at the expense of the no HNO with
415 LNO and the no HNO no LNO categories.

416 Further investigation was undertaken to determine the spatial relationship between
417 poly(dA)/G4 enriched sequences and nucleosome distribution. To achieve this, strand
418 separated poly(dA) enriched sequences and predicted G4s were overlapped with the MNase-
419 seq data [50] (Methods). The plot obtained after the intersections revealed that the LNO region
420 was positioned between the plus and minus poly(dA)/G4 summits (Figure 6D). Additionally,
421 the HNO peaks were found to exhibit overlap with the poly(dA) and G4 peaks and to
422 surround the origins (Figure 6D). Furthermore, these intersections permitted the estimation
423 of the averaged distance of the above-mentioned elements from the centre of origins. The
424 LNO region extended to approximately ±100 to 150 nt from the centre, thereby defining an
425 LNO area of approximately 200 to 300 bp in the centre of origin. The two HNO regions were
426 located between 150 and 500 nt upstream and downstream from the centre of origin (Figure
427 6D).



428

429 **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 [50] 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 [50] 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) [50] 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 [50] (blue dotted line) around centred origins and intergenic shuffled controls (± 2 kb). The G4s were predicted by the G4Hunter application [51] 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).

The origins colocalized with a subset of DRIP-seq enrichment sites.

DRIP-seq is a widely used technique for genome-wide mapping of RNA:DNA hybrids. 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 in *T. brucei* BSF cells, revealing 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 RNA:DNA hybrids was identified around the origins, with two distinct summits located close to the origin centre (Figure 7A). An 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 RNA:DNA hybrids (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, the 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 the DRIP-seq enrichment (Figure 7B); a statistically significant correlation was observed (odds ratio = 4.17; *P*-value $2.87E^{-142}$), suggesting a link between RNA:DNA hybrid formation and origin activity. As DRIP-seq was only performed on BSF cells [54], we intersected this set of data with PCF and BSF origins separately and found that the distribution of RNA:DNA hybrids around BSF origins was 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 SNS-seq and DRIP-seq overlap 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

476 intersections also revealed that only a small proportion of the previously identified RNA:DNA
477 hybrids (R-loops) [54] overlapped with the merged origins (1.7%), PCF origins (0.7%), or BSF
478 origins (1.3%).

479 Subsequently, the predicted G4 structures and poly(dA) enriched sequences from the plus
480 and minus strands were superimposed with the DRIP-seq dataset [54] at the centres of origins
481 and shuffled controls (Figure 7C). This intersection allowed us to estimate the relative
482 positions of these elements and their averaged distance from the centre of origins. The two
483 distinct summits of R-loops were observed to be situated, on average, between the centre
484 and 200 nt upstream and downstream from the centre of origins (Figure 7C).

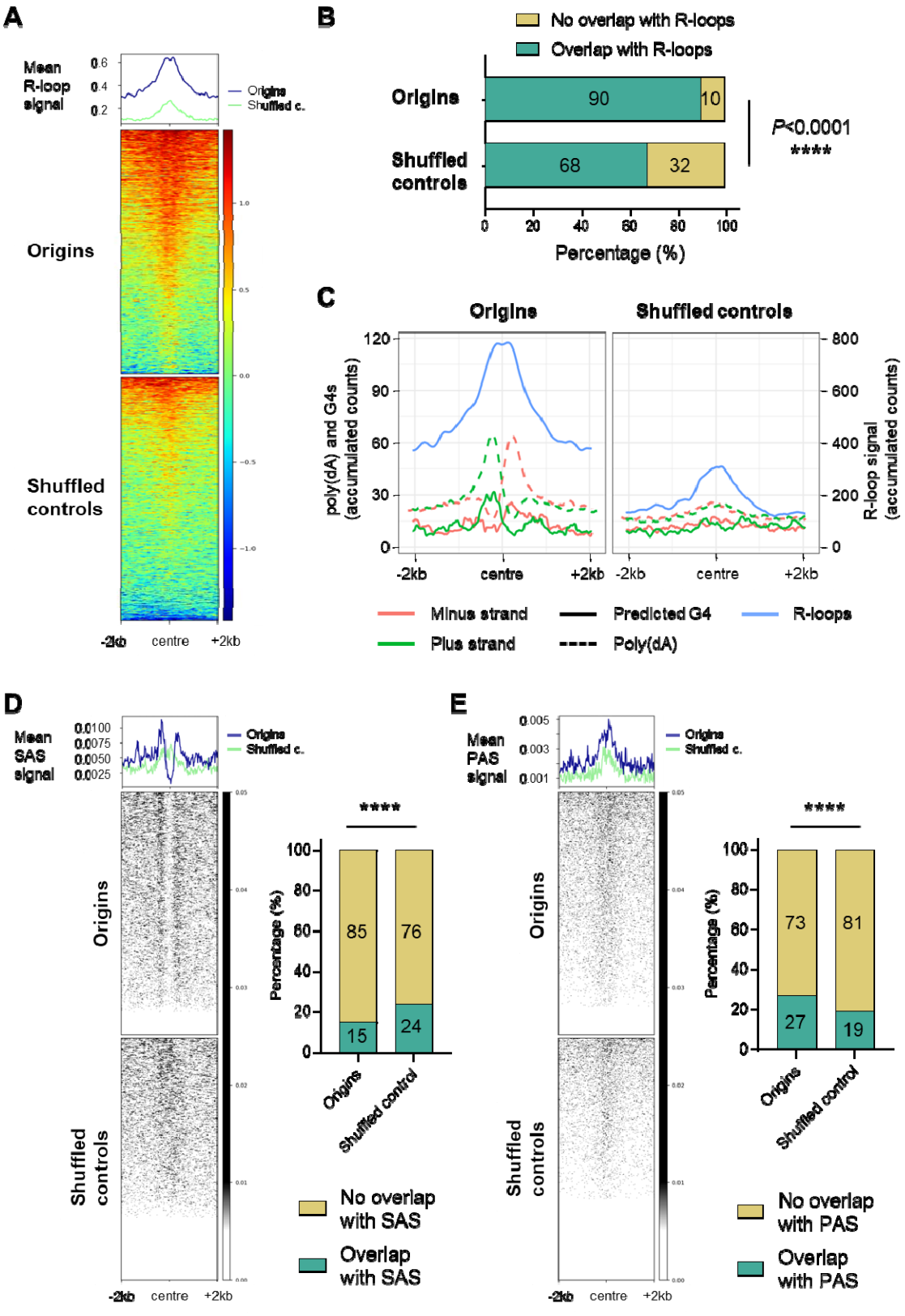


Figure 7. The distribution of DRIP-seq enrichment, splice acceptor sites (SAS) and polyadenylation sites (PAS) around the origins.

488 A) The profile plots and heatmaps show the distribution of R-loops [54] around the centred origins (blue line) and
489 the intergenic shuffled controls (green line) (± 2 kb). Mean DRIP-seq signal presents average DRIP-seq signal [54]
490 per 20 bp window. B) The proportion of origins and shuffled controls that overlap with R-loops (intersection
491 without window). The *P*-values were calculated using Fisher's exact (two-sided) test, which involved a comparison
492 of the numbers of indicated categories for each pair of datasets (**** - $P<0.0001$). C) The profile plots illustrate the
493 distribution of the stranded poly(dA) sequence (AAAA) (dashed line), predicted G4s (solid line) and R-loops [54]
494 (solid blue line) around centred origins and shuffled controls (± 2 kb). The plus strand is represented in green, and
495 the minus strand in pink. The 13,409 G4s were predicted by the G4Hunter application [51] in the Tb Lister 427-2018
496 reference genome. The values on the plots present the accumulated counts of the AAAA, G4s and R-loop signals
497 per position within a 100 bp window (50 bp up- and downstream from the position). D) Left panel: the profile
498 plots and heatmaps show the distribution of transcription splice acceptor sites (SAS) around the centred origins
499 (blue line) and the intergenic shuffled controls (green line) (± 2 kb). Intersection was performed in the Tbb
500 TREU927 reference genome for the 11 megabase chromosomes. Right panel: the proportion of origins and
501 shuffled controls that overlap with SAS (intersection without window). The *P*-value was calculated using Fisher's
502 exact (two-sided) test, which involved a comparison of the numbers of indicated categories for each pair of
503 datasets (**** - $P<0.0001$). E) Left panel: the profile plots and heatmaps show the distribution of transcription
504 polyadenylation sites (PAS) around the centred origins (blue line) and the intergenic shuffled controls (green line)
505 (± 2 kb). Intersection was performed as in Figure 7D. Right panel: the proportion of origins and shuffled controls
506 that overlap with PAS. The *P*-values were calculated using Fisher's exact (two-sided) test, which involved a
507 comparison of the numbers of indicated categories for each pair of datasets (**** - $P<0.0001$).

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

510 We compared the replication origins identified by stranded SNS-seq with the data obtained
511 by Marker frequency analysis coupled with deep sequencing (MFA-seq) [42,43]. MFA-seq
512 compares read coverage between replicating and non-replicating cells and identifies regions
513 with increased DNA coverage in S phase, representing replicated regions. In *T. brucei*, MFA-
514 seq identified broad regions of a few hundred kilobases [42,43], whereas the origins detected by
515 stranded SNS-seq are about 150 bp long. We used six MFA-seq datasets [42,43] (Methods) and
516 we presented core chromosomal overview of the MFA-seq regions [42,43], TbORC1/CDC6
517 binding sites [42] and the stranded SNS-seq origins on the 11 core chromosomes of *T. brucei*
518 (Supplementary Figure 8A). We found that 28-42% of the origins identified by stranded SNS-
519 seq overlapped with early and 43-55% overlapped with late S-phase MFA-seq replicated
520 regions (Supplementary Figure 8B). We also reanalysed the overlap between TbORC1/CDC6
521 binding sites [42] and MFA-seq regions [42,43]. Our findings indicated that this overlap is in a
522 similar range as the overlap between SNS-seq origins and MFA-seq replicated regions
523 (Supplementary Figure 8B).

524 Several proteins have been identified as potential subunits of the origin recognition complex
525 (ORC) in *T. brucei*. One of these is TbORC1/CDC6, which complemented the yeast protein
526 Cdc6, but not Orc1, in a phenotypic complementation assay [55]. We intersected the origins
527 detected by stranded SNS-seq with the previously detected binding sites of the
528 TbORC1/CDC6-12Myc tagged protein [42] (Methods). For this intersection, we used the data
529 from TriTrypDB that contains 6482 ORC1/CDC6 binding sites, we filtered this dataset at score
530 22 to obtain 990 binding sites which is close to the 953 TbORC1/CDC6 binding sites
531 described previously [42] (Methods). The intersection revealed that TbORC1/CDC6 levels are

532 lowest at the centre of origins (Supplementary Figure 8C). We observed that 12.2% of SNS-
533 seq origins overlap with 24.2% of TbORC1/CDC6 binding sites within a ±2 kb window, 17% of
534 SNS-seq origins overlap with 24.2% of ORC1/CDC6 binding sites within a ±3 kb window and
535 26.9% of SNS-seq origins overlap with 53.4% of ORC1/CDC6 binding sites in ±5 kb window.
536 Therefore, we identified a depletion of TbORC1/CDC6 binding sites at the centre of the
537 origins, accompanied by a slight accumulation of ORC1/CDC6 binding sites 5 kb upstream
538 and downstream of the origins (Supplementary Figure 8C).

539 **Spatial coordination between the activity of the origins and transcription.**

540 DNA replication in trypanosomatids occurs in a particularly challenging environment, as
541 most of their genomes are constitutively transcribed. Arrays of genes oriented in the same
542 direction are transcribed by RNA polymerase II into polycistronic transcription units (PTUs),
543 which can contain hundreds of genes (reviewed in [40]). The primary transcript is processed
544 into individual mRNAs by trans-splicing of a capped 'spliced leader' at the splice acceptor
545 site (SAS) and by polyadenylation at the polyadenylation sites (PAS) (reviewed in [40]). To gain
546 insight into the relationship between origins and SAS and PAS, we intersected our dataset
547 with the unified SAS and PAS datasets for the core chromosomes (Methods). The intersection
548 of origins with SAS showed that the centre of origins was depleted of SAS, while some
549 origins had SAS flanking the centre and the SAS signal was enriched ±200-500 nt from the
550 centre of origins (Figure 7D, left panel). Only 15% of the origins overlapped with SAS
551 compared to 24% of the shuffled controls, and this difference was statistically significant
552 (Figure 7D, right panel). The intersection of the origins with PAS revealed an accumulation of
553 PAS near the origin centres. A similar, though less pronounced, accumulation was also
554 observed in the shuffled controls (Figure 7E, left panel). Quantification showed that 27% of
555 origins overlapped with PAS compared to 19% of shuffled controls and this difference was
556 statistically significant (Figure 7E, right panel).

557 PTUs are separated by strand-switching regions (SSRs) and head-to-tail (HT) regions. SSRs
558 are classified as either divergent (dSSRs) or convergent (cSSRs). dSSRs and cSSRs serve as
559 transcription start (TSS) and termination sites (TTS), respectively (reviewed in [40]). Our results
560 showed that the origins are predominantly located in the intergenic regions (Figure 2C and
561 2D). The metaplots revealed that the TSS, TTS and HT regions [56] were not enriched with
562 origins identified by stranded SNS-seq (Supplementary Figure 8D). Only 3.7% of origins are
563 located within 30% of TSS, TTS and HT regions. Therefore, most origins are localized in
564 actively transcribed regions, which could lead to collisions between DNA replication and the
565 transcription machinery. This spatial coincidence implies that transcription and replication
566 must occur in a highly ordered and cooperative manner in *T. brucei*.

567 **DISCUSSION**

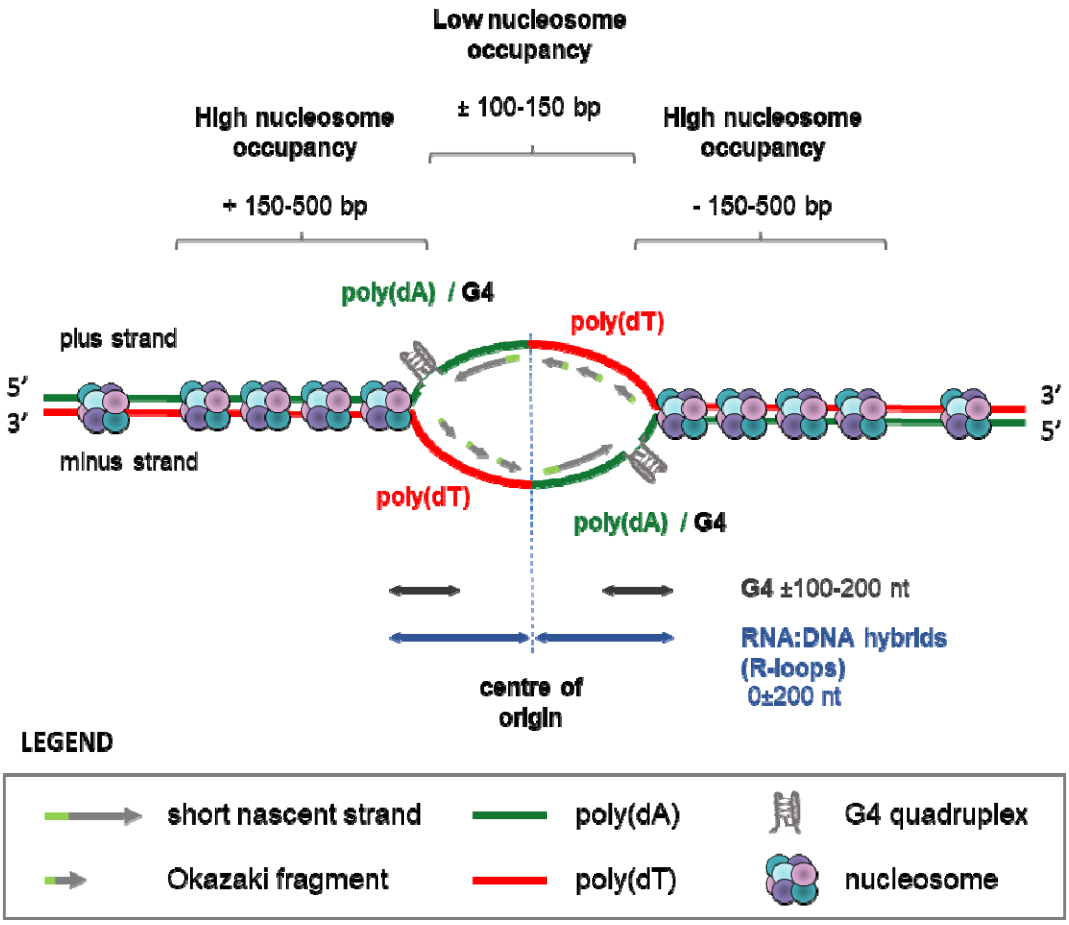
568 We used stranded SNS-seq to identify a set of active origins in the unicellular eukaryote, *T.*
569 *brucei*. Our methodology is comparable to previously published origin-derived single-
570 stranded DNA sequencing (Ori-SSDS) [17] but includes few differences. The stranded library

571 preparation process was different in these two approaches, and the SNS-depleted control
572 was included in the origin mapping process as a background control. The key benefit of
573 stranded over conventional (non-stranded) SNS-seq is the ability to identify origins with
574 greater specificity by eliminating peaks that do not align with the proper orientation of SNS
575 of an active origin. In addition, this method maps the origins between two divergent SNS
576 peaks, allowing the determination of the centre of origins or replication initiation sites, which
577 was essential for the discovery of the positions and orientations of the elements involved in
578 the topology of origins. Origins mapped between two divergent SNS peaks showed the
579 origin-specific nucleotide distribution around the centre (Figure 4A, Supplementary Figure
580 9A). On the other hand, if we apply conventional SNS-seq origin mapping to our data and
581 map origins in the middle of peaks without knowing their orientation, the origin-specific
582 nucleotide distribution is lost (Supplementary Figure 9B). Therefore, the primary benefits of
583 our approach include the ability to centre origins between two divergent SNS peaks and to
584 eliminate false positive peaks.

585 Using this stringent origin selection procedure, we have selected and analysed a subset of all
586 origins. Some weak or inefficient origins may be missed due to low activity, resulting in low
587 read coverage. It is therefore very likely that the number of origins of this parasite is higher
588 than the number reported here. We also examined the origins separately in the two life
589 stages demonstrating that the origin structures in the PCF and BSF cells were equivalent and
590 justifying the merging of these two sets of origins into one. Under this stringent origin
591 mapping condition, we found that 238 origins are in common (Supplementary Figure 2C).
592 However, we decided not to compare the PCF and BSF sets of origins in more detail because
593 comparing non-exhaustively mapped origins in these two life cycle stages could be
594 misleading rather than indicative of true differences between the two cell types.

595 Our analysis revealed no consensus sequence associated with the origins, which is consistent
596 with the findings of most other studies of eukaryotic origins, including trypanosomatids
597 [22,27,42,43]. Nevertheless, the genomic sites that function as origins are not random sequences.
598 Thanks to the stranded SNS-seq approach, we were able to determine the positions and
599 strand-specificity of some genetic elements and to present a comprehensive model of an
600 origin (Figure 8). The results of our study showed that DNA replication starts at genomic sites
601 where adenine and thymine are overrepresented and arranged in a specific pattern. On the
602 plus strand, poly(dA) enriched sequences were situated upstream, while poly(dT) enriched
603 sequences were located downstream of the origin's centre (Figure 4B and Figure 8). The G4s
604 near the centre of origins were embedded in poly(dA) rich sequences and the summits of G4
605 peaks covered 100-200 nt up and downstream from the centres (Figure 5D and Figure 8).
606 Our model shows that SNS molecules are synthesised using poly(dA)/G4-rich
607 complementary strands as templates (Figure 8). In our study the enriched molecules
608 detected in divergent peaks do not contain G4s. Rather, they are complementary to the
609 strands enriched with G4s. Therefore, the observed enrichment of G4s around origins cannot
610 be an artefact of the enzymatic treatments, as previously suggested [35]. The centres of origin

611 were identified as LNO regions, which had an average length of 200–300 nt and were flanked
612 by HNO regions located from ±150 to ±500 bp from the centre (Figure 6D, Figure 8).
613 RNA:DNA hybrid enrichment revealed two distinct summits, one upstream and one
614 downstream of the centre, with a range spanning ±200 nucleotides (Figure 7A, 7C and Figure
615 8).



616

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

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

627 Our analysis indicated that G-rich sequences near origins tend to form G4 structures (Figure
628 5A-C, Supplementary Figure 5B-C). The detection of more G4s under PDS-stabilized
629 conditions compared to physiological ones suggests that there is an important potential for
630 G4s formation near origins. G4 formation seems to be very transient and therefore not easy

631 to detect unless stabilized by drug. It seems that G4 structures near origins undergo rapid
632 changes, either forming or dissolving in response to the replication dynamics. We therefore
633 propose that the formation of G4s plays a regulatory role in the activation of replication
634 origins. Some of the recent publications suggested that G4s may serve as replication fork
635 barriers and play a role in transient replication fork stalling [21,57]. The enrichment of G4
636 structures on strands acting as templates for leading strand synthesis implies that the
637 formation of G4 structures could potentially pose or arrest replication fork progression.

638 Our analysis disclosed for the first time a strong link between RNA:DNA hybrid formation
639 and DNA replication initiation. Given that RNA-primed nascent strands naturally form
640 RNA:DNA hybrids during replication initiation, the enrichment of RNA:DNA hybrids near
641 origins is both expected and biologically relevant. We found that a small proportion of
642 previously reported RNA:DNA hybrids (1.7%) overlap with SNS-seq origins and we suggest
643 that these RNA:DNA hybrids represent the SNS-derived RNA:DNA hybrids et the DNA
644 replication initiation sites. The S9.6 antibody used for immunoprecipitation in DRIP-seq
645 experiments exhibits a high affinity for RNA:DNA hybrids longer than 6 bp [58]. The RNA:DNA
646 hybrids are formed at the replication initiation sites by RNA priming of the SNS and Okazaki
647 fragments and S9.6 antibody can immunoprecipitate these RNA:DNA hybrids. However,
648 Okazaki fragments are not exclusive markers of origins as they form RNA:DNA hybrids
649 throughout the genome, whereas RNA:DNA hybrids formed by RNA-primed SNS are specific
650 markers of origins. Therefore, we suggest that the summits of RNA:DNA hybrids near the
651 origins reflect the enrichment of RNA-primed SNS molecules. We also consider that the
652 overlap between the two distinct techniques, DRIP-seq and stranded SNS-seq, provides
653 compelling validation of our origin mapping approach.

654 The elements presented in our model were not found in all origins, as revealed by the
655 quantification performed in this study. This may be caused by the fact that the datasets used
656 for the intersections were obtained from four different strains, rather than from just one.
657 Different strains are likely to activate different sets of origins, which could introduce bias
658 during the intersection process. Another possible source of bias may be the varying quality
659 of the genome assemblies, since this study used three different reference genomic
660 sequences (Methods). This is clearly demonstrated by the detection of different numbers of
661 origins in different reference genomic sequences using the same stranded SNS-seq data
662 (Methods). An alternative hypothesis to explain the absence of certain structures in some
663 origins is that they form transiently, only during origin activation. Therefore, it can be
664 postulated that the presence of most of the elements, in addition to the chromatin state and
665 nuclear organisation of origins, defines the moment of origin activation.

666 Different origin mapping methods often demonstrate limited overlap due to biases related to
667 the chosen methodology [59,60]. SNS-seq and MFA-seq are two complementary techniques
668 that provide information on different aspects of DNA replication. SNS-seq is a widely used
669 method of high-resolution origin mapping, whereas MFA-seq identifies replicated genomic
670 regions in synchronized cells by comparing the number of DNA copies in the S phase and

671 other phases of the cell cycle [60,61]. The broad replicated regions (0.1–0.5 Mbp) identified by
672 MFA-seq in *T. brucei* [42] are likely to contain multiple origins, rather than just one. In this
673 regard, we disagree with the interpretation of McCulloch's group, who postulated that there
674 is a single origin per broad peak [42]. Our analysis shows that up to 28-55% of the origins
675 detected by stranded SNS-seq locate within MFA-seq regions, providing complementary
676 information on the position of origins within broad replicated regions. It should also be
677 noted that the experimental context was different: MFA-seq relies on cell sorting, while
678 stranded SNS-seq is performed on asynchronous cell populations, enabling the identification
679 of origins active throughout S phase. A similar result was found when comparing
680 conventional SNS-seq and MFA-seq in another trypanosomatid, *Leishmania major* [22].

681 We found a minority of TbORC1/CDC6 binding sites near origins detected by SNS-seq and
682 this was comparable with the previously reported ~4.4% overlap of this protein with MFA-
683 seq data (reviewed in [62]). Comparable results were also obtained in *T. cruzi* [27]. DNAscent
684 nanopore sequencing revealed 4,287 origins in *T. cruzi*. Of these, 3,782 (88.2%) did not match
685 the 851 TcORC1/CDC6-Ty1 binding sites within a ±3 kb window [27]. Therefore, only 11.7% of
686 origins of *T. cruzi* overlapped with TcORC1/CDC6-Ty1 binding sites within a ±3 kb window, a
687 proportion similar to what we observed in *T. brucei* (6.2%). The origins detected by
688 conventional SNS-seq in *Plasmodium falciparum* revealed 63% of overlap with initiation
689 events detected by nanopore sequencing and 13.6% with $PfORC_{1-2}$ binding sites [36]. Another
690 group found no significant correlation between origins detected by DNAscent nanopore
691 sequencing and PfORC1 binding sites within the same organism [63]. Furthermore, there is low
692 overlap between origins mapped using different techniques and ORC-binding sites in human
693 cells [59]. We did not detect any accumulation of origins detected by stranded SNS-seq within
694 SSRs. This differs from the previously reported accumulation of TbORC1/CDC6 binding sites
695 at divergent SSRs in the Tb927 genome [42]. The reason for the low correlation between origins
696 detected by different methods and ORC binding sites remains unknown, and further studies
697 are necessary to answer this complex question.

698 It is important to note that the topology of origins of *T. brucei* is very similar to that
699 previously described for other eukaryotes. The origins in mouse cells have been
700 characterised as fragile sites, defined by long poly (dA:dT) tracks that are nucleosome free
701 and devoid of the single-strand DNA-protecting protein RPA [13]. The same position and
702 orientation of predicted G4 motifs, peaking at 240-300 base pairs from replication start sites,
703 was previously identified in two drosophila cell lines using conventional SNS-seq [21], as was
704 found here for *T. brucei*. Furthermore, the organisation of nucleosomes around the origins of
705 *T. brucei* is identical to that observed in yeast and higher eukaryotes [12,28–32,64,65]. In the case of
706 *L. major* the origins identified using conventional SNS-seq were also located in genomic
707 regions with lower nucleosome occupancy [22]. In addition, a significant correlation between *L.*
708 *major* origins and predicted G4 motifs was observed, with 74% of origins containing at least
709 one such motif [22]. These similarities provide strong support for the reliability of our origin
710 identification in *T. brucei* and suggest that the structure of origins may be a conserved

711 feature in all eukaryotes. However, further studies are needed to gain a deeper insight into
712 the evolution of DNA replication origins.

713 **METHODS**

714 **Cell culture**

715 Procyclic forms (PCF) of *T. brucei*, strain Lister 427 wild type strain were grown at 27°C in
716 SDM-79 (PAA Laboratories) supplemented with heat inactivated 10% FBS (Gibco) and
717 7µg.mL^{-1} hemin [66]. Bloodstream forms (BSF) of *T. brucei* Lister 427 strain were cultivated at
718 37°C, 5% CO in a modified Iscove's culture medium (Gibco) complemented according to
719 Hirumi *et al.* [67], with 10% (v/v) heat-inactivated foetal bovine serum (Gibco), 36 mM sodium
720 bicarbonate (Sigma), 136 µg.mL^{-1} hypoxanthine (Sigma), 39 µg.mL^{-1} thymidine (Sigma), 110
721 µg.mL^{-1} Na-pyruvate (Sigma), 28 µg.mL^{-1} bathocuproine (Sigma), 0.25 mM β-
722 mercaptoethanol, 2 mM L-cysteine.

723 **SNS isolation and purification**

724 SNS were isolated and purified following a previously published protocol [16] with few
725 modifications. Between 2 - 2.5x10^{9} cells were used as starting material for SNS isolation. Total
726 DNA was isolated from three biological replicates of asynchronous PCF and BSF cells in the
727 exponential phase of growth (3-5x10^{6} and 7-9x10^{5} cells.mL^{-1} for PCF and BSF respectively).
728 Cells were precipitated and incubated in 0.5mL of 20 mM Tris-Cl, pH 7.0, 50mM EDTA, 1%
729 N-Lauroylsarcosine, 0.5% Triton X-100 and 2mg.mL^{-1} proteinase K and incubated at 45°C
730 for 24 hours. Fresh solution of proteinase K was added after 24 hours and digested for 16
731 hours more. DNA was purified with phenol/chloroform/isoamyl alcohol and precipitated in
732 0.3 M Na-acetate (pH 5.3) and absolute ethanol. DNA was dissolved gently in TEN20 buffer
733 (10 mM Tris-Cl, 50 mM EDTA, 20 mM NaCl, pH 7.0) with 100 U RNasin (Promega). DNA was
734 heat-denatured (95°C/5 min) and loaded onto 5 mL of neutral 5%–20% sucrose gradients
735 prepared in TEN20 buffer in ultra-clear tubes (Beckman Coulter). Size separation of nucleic
736 acids was done by ultra-centrifugation in SW50 rotor for 7 hours at 45 krpm at +4°C.
737 Fractions of 200 µL from the sucrose gradient were collected and checked by agarose gel
738 electrophoresis to assess the quality of nucleic acid separation (Supplementary Figure 10A).
739 Fractions from 0.5-2.5 x10^{3} nt were pooled together as SNS enriched fractions. Half of the
740 SNS pool was treated with 10 µg of cocktail of RNase, DNase free (Roche, Cat Nb
741 11119915001) for 16h at 37°C, to become the SNS-depleted control. SNS-enriched samples,
742 together with its SNS-depleted controls, were treated 3-4 times with 50 U of T4 PNK (Thermo
743 Fisher Scientific) and 100 U of λ-exonuclease (Thermo Fisher Scientific). Each treatment was
744 done after heat denaturation of DNA (95°C/10 min) and simultaneously on samples and
745 controls. After denaturation, samples and controls were immediately placed on ice. The
746 enzymatic reactions were assembled on ice, and then the enzymatic treatment with T4 PNK
747 and λ-exonuclease was done at 37°C. After enzymatic treatment and prior to library
748 preparation, residual RNA was removed with 5 µg RNase, DNase-free (Roche Cat Nb
749 11119915001), isolated DNA was sonicated with Bioruptor Pico sonication device

750 (Diagenode) and purified with magnetic beads (CleanNGS, Proteigen). The quantity of
751 purified material was estimated by fluorometer DS 11 FX (DeNovix) using Qubit ssDNA Assay
752 kit (Invitrogen).

753 It is important to note that we used SNS enrichment conditions and enzymatic treatments as
754 described in previous reports [16,68]. We enriched the SNS by size on a sucrose gradient and
755 then treated this SNS-enriched fraction with high amounts of λ-exonuclease. A previous
756 study has shown that under these conditions, complete digestion of DNA containing G4-rich
757 sequences is achieved [20]. The efficiency of the T4 PNK and λ-exonuclease enzymes was tested
758 on both sheared gDNA and on RNA. The majority of DNA was digested after one treatment
759 with these two enzymes (Supplementary Figure 10B). RNA molecules treated under the same
760 conditions were preserved (Supplementary Figure 10C), suggesting that SNS would also be
761 preserved. The quantity of DNA molecules in SNS containing fractions was systematically
762 followed by qPCR after each enzymatic treatment and it was estimated that the amount of
763 DNA in starting material was reduced around 100 times after the first treatment, and around
764 10 times more after each subsequent treatment with these two enzymes.

765 **Preparation of stranded libraries for paired end sequencing**

766 Single-stranded SNS libraries were prepared using the Accel-NGS 1S Plus DNA Library Kit
767 and 1S Plus Dual Indexing Kit for Illumina (Swift Biosciences), following the manufacturer's
768 protocol. A 3'end specific tail is preserving the strand orientation of the ssDNA fragments
769 during library preparation, resulting in a second strand library, where the plus-strand
770 fragments are represented by F1R2 read pairs and the minus-strand fragment by F2R1 read
771 pairs. The quality of the libraries was checked by capillary electrophoresis on Bioanalyzer
772 2100 (Agilent) using Agilent High Sensitivity DNA Kit (Agilent Technologies). The quantity of
773 libraries was estimated by fluorometry using dsDNA High Sensitivity Assay Kit on a DS-11
774 FX+ (DeNovix).

775 **Paired end sequencing (2x150)**

776 The libraries were validated for quality and quantity using Fragment Analyzer and qPCR
777 techniques. qPCR was performed using KAPA Library Quantification Kits - Complete kit
778 (Universal) (Roche, ref.7960140001) and Light Cycler 480 machine (Roche). The libraries were
779 pooled equimolarly and sequenced on the Illumina NovaSeq 6000 using a NovaSeq Reagent
780 Kit for 300 cycles (Illumina) following manufacturer's instructions. Libraries were sequenced
781 on one lane of a SP flow-cell in paired-end 150 nt. Sequencing was performed by the
782 GenomiX (MGX) Core Facility, Montpellier, France. Demultiplexing and FASTQ files production
783 was done using the Illumina software bcl2fastq (RRID:SCR_015058, v2.20.0.422).

784 **SNS-seq data analysis**

785 Quality control of the FASTQ files was performed using FastQC (RRID:SCR_014583, v0.11.9)
786 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The sequencing reads were tail

787 trimmed to remove the 3'end specific tail introduced during library preparation using
788 cutadapt [69] with the following parameters (-u -10 -u 10 -U -10 -U 10). Afterwards the SNS-
789 seq sequencing reads were aligned against the entire *T. brucei* Lister 427-2018 reference
790 genome release 55 from TriTrypDB (RRID:SCR_007043)
791 (https://tritrypdb.org/common/downloads/release-55/TbruceiLister427_2018, last accessed
792 June 24th, 2025) with bowtie2 (RRID:SCR_016368 , v2.4.10) [70]. The alignment output bam files
793 were converted, sorted, and indexed using samtools [71]. Duplicated reads were removed using
794 picard (RRID:SCR_006525, v2.23.5, "Picard Toolkit." 2019. Broad Institute, GitHub Repository.
795 https://broadinstitute.github.io/picard/, last accessed June 24th, 2025; Broad Institute).
796 Mapped deduplicated reads pairs were then separated with samtools (RRID:SCR_002105,
797 v1.13) *view* into minus-strand and plus-strand read pairs via their read orientation (minus-
798 strand: F2R1 and plus-strand: F1R2). Peak calling was then performed with macs2 (v.2.2.7.1)
799 (using default parameters plus -p 5e-2 -s 130 -m 10 30 --gsize 2.5e7) for plus and minus-
800 strand reads separately with the SDC as control.

801 **Origin selection**

802 Origins were selected after the peak calling step, during the peak filtering step. The called
803 peaks were filtered using the bedtools suite (RRID:SCR_006646, v.2.30.0) [72] as described
804 below. In this peak filtering step, three criteria were defined to select the origins. The first
805 criterion was that two SNS peaks must be oriented in a way that reflects the topology of the
806 SNS strands within an active origin: the upstream peak must be on the minus strand and
807 followed by the downstream peak on the plus strand. The second criterion was to set a
808 maximal inner to inner border distance between two divergent minus and plus peak pair. The
809 distances between the start points of the two divergent leading strands are not known. We
810 conducted tests on several inner border-to-border distances, ranging from 0.5 to 2.5 kb.
811 Higher numbers of origins were found with larger distances. However, we finally applied a
812 maximum distance of 0.5 kb, presuming that this corresponds more closely to the
813 physiological distance between two divergent SNS molecules. The third criterion was
814 introduced to prevent complete overlap of two divergent peaks, as this can only occur if the
815 two SNSs elongate at the same rate in both directions, which is an unlikely scenario. Indeed,
816 the synthesis of the Okazaki fragment and ligation to the 5' of the SNS requires several
817 enzymatic processes, and we can assume that the elongation of SNS fibre is slower at the 5'
818 end. Of note, complete overlap of minus and plus peaks was rarely observed (mean 6.5%
819 (stdev 6.5%, min 0.8%, max 15.8%)). To define the maximum overlap between the minus and
820 plus SNS peaks, a series of tests were conducted, varying the percentage of overlap from
821 40% to 90%. This resulted in a similar number of origins, and ultimately, 50% was selected as
822 the optimal value. Therefore, firstly, minus-strand and plus-strand peaks with an overlap of
823 over 50% were removed with bedtools *intersect*. Secondly peak pairs on different strands
824 with a maximal distance of 0.5 kb were selected using bedtools *windows*. Finally, peak pairs
825 were filtered for the correct orientation (minus-strand followed by plus-strand) with bedtools
826 *closest* and the origin was defined as the genomic region between the two inner borders of

827 two divergent peaks. The overlap of origins between replicates and cell types as well as other
828 genomic features was analysed with bedtools *windows* and bedtools *intersect*. Replicates
829 were merged with bedtools *merge*.

830 For **visual inspection of genomic features** Integrative genomics viewer (IGV)
831 (RRID:SCR_011793) [73] was used. Violin plots were generated with the GraphPad Prism 8.3.0
832 software (RRID:SCR_002798, GraphPad Software Inc., La Jolla, CA, USA).

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

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

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

841 **Read sub-sampling** was performed to be able to compare origin number in BSF and PCF.
842 The mapped reads from all three PCF samples and control and BSF samples and control,
843 respectively, were merged using Samtools *merge*. The PCF combined sample was sub-
844 samples with Samtools *view* (--subsample 0.69 --subsample-seed [0-14]) and the merged
845 control with Samtools *view* (--subsample 0.57 --subsample-seed [0-14]) 15 times to obtain a
846 set of mapped reads comparable to the BSF combined sample. Origin detection was applied
847 as described above to the combined BSF and the 15 subsamples combined PCF samples.

848 For **visualization of the respective** localization of two or more datasets the the Deeptools
849 suite (RRID:SCR_016366, v3.5.1) [74] and the ggplot2 R package was used. The open source
850 ggplot2 R package (RRID:SCR_014601), freely available from the Comprehensive R Archive
851 Network (CRAN) repository (http://www.r-project.org, last accessed June 24th, 2025) was used
852 to generate composite plots. In Deeptools bed and bigwig files were used to generate a
853 matrix file with *ComputeMatrix* and *PlotHeatmap* or *PlotProfil* were used to generate
854 visualizations from the matrix.

855 **Motive enrichments** were tested over the origin centres extended 250 bp up and
856 downstream with the MEME suite motive discovery (RRID:SCR_001783) [75] as well as with
857 homer (RRID:SCR_010881) *findMotifs* [76]. A multiple sequence alignment and the analysis of
858 nucleotide patterns were built with the WebLogo (RRID:SCR_010236) online application [47].

859 **Nucleotide composition** around the origin centres was calculated as follows. The *T brucei*
860 genome was split into 20 bp windows using bedtools *window*. The number of A, T, G and C
861 nucleotides for each window were obtained with bedtools *nuc* and used to calculate the
862 nucleotide percentage per 20 bp window. The resulting bedgraph files were transformed into
863 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 [50], DRIP-seq [54], G4-seq [49], 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/).

Transcription start sites (TSS), transcription termination sites (TTS) and HT regions. A list of TSS, TTS and head-to-tail (HT) regions for the Tbb TREU927 reference genome was obtained from Kim et al. [56].

MNase-seq. The MNase-seq data [50] 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 [50]) 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 [50] 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

902 corresponds to the average value for background regions (2000-1000 bp upstream of the
903 origin centres and 1000-2000 bp downstream of origin centres).

904 **Drip-seq.** The Drip-seq dataset was downloaded from ENA (accession number PRJEB21868)
905 [54] as fastq files. The re-analysis was done as described [54]. The published DRIP-seq analysis
906 was performed in Tb Lister 427 reference genome. We re-analysed the data in the Tb Lister
907 427 and the Tb Lister 427-2018 reference genomes. The intersection of DRIP-seq data with
908 the stranded SNS-seq origins was done in the Tb Lister 427-2018 reference genome.

909 **G4-seq.** The G4-seq [49] dataset was downloaded from GEO (accession number GSE110582) as
910 bed and bedgraph files separated for the plus and minus strand. G4-seq was performed on *T.*
911 *brucei* EATR01125 strain and analysed in the Tbb TREU927 reference genome [49] to be able to
912 directly compare the data with the SNS-seq dataset we analysed the SNS-deq data in the
913 Tbb TREU927 reference genome.

914 **G4 prediction** in the *T. brucei* genome were done with G4Hunter [51] with the following
915 settings, window size (-w 25) and a threshold of 1.57. Prediction of G4 structures in the Tb
916 427-2018 reference genome and in the Tb Lister 427 reference genome were performed, to
917 be able to compare G4 structures to MNase-seq (in Tb Lister 427 reference genome) and
918 DRIP-seq (in Tb Lister 427-2018 reference genome) data as G4-seq output files were
919 available only for the Tbb TREU927 reference genome [49].

920 **MFA-seq.** We intersected six MFA-seq datasets. Two were obtained directly from the authors
921 [42] and four were downloaded from ENA (accession number PRJEB11437) [43], all in fastq file
922 format. The analysis of the fastq files was performed as described previously [42,43]. The original
923 published MFA-seq analysis was performed in *T. brucei* Tbb TREU927 reference genome. We
924 re-analyzed the data in the Tbb TREU927 reference genome and in the Tb Lister 427-2018
925 reference genomes to be able to directly compare them to the stranded SNS-seq data.

926 **ORC1/CDC6 ChIP-on-chip.** The Orc1/CDC6 ChIP-on-Chip dataset [42] was downloaded from
927 TriTrypDB (as ratios_peaks) for each element in the *T. brucei* Tbb TREU927 reference genome
928 in form of bed files. The dataset available on TriTrypDB differs from the one published in the
929 manuscript of Tiengwe et al. [42] but the originally published dataset is not available anymore
930 (personal communication Richard McCulloch). The TriTrypDB ORC1/CDC6 dataset contains
931 6543 binding sites whereas the dataset published originally contained 953 binding sites [42], to
932 obtain a number comparable to the published number of ORC1/CDC6 binding sites we
933 filtered the TritrypBD data set at a score of 22 which resulted in a set of 990 ORC1/CDC6
934 binding sites. To be able to directly compare them to the SNS-seq we analysed the SNS-seq
935 data in the Tbb TREU927 reference genome.

936 **DNA molecular combing**

937 Asynchronous cell populations of *T. brucei* were grown to the density of $4x10^6$ and $7x10^5$
938 cells.mL^{-1} for PCF and BSF, respectively. Cells were sequentially labelled with two modified
939 nucleosides, firstly with 300 µM iodo-deoxyuridine (IdU, Sigma) for 20 minutes and then 20

940 minutes with 300 µM chloro-deoxyuridine (CldU, Sigma) without intermediate washing.
941 After labelling, the cells were immediately placed on ice to stop DNA replication and
942 extensively washed with ice-cold 1xPBS. Around $1x10^8$ cells were blocked in 1% low melting
943 agarose plugs.

944 DNA molecular combing was done as described previously [41]. Protein-free plugs with DNA
945 were extensively washed in TEN20 buffer (10 mM Tris-Cl, 50 mM EDTA, 20 mM NaCl, pH 7.0)
946 and then stained with 0.5 µM YOYO-1 fluorescent dye (Molecular Probes) for 1 h at 37°C.
947 After washing from the excess YOYO-1 day, the 200 µL of TEN20 buffer was added to the
948 plugs, melted at 65 °C for 15 minutes, slowly cooled down to 42 °C and treated with 10 U
949 of β-agarase (Merck) for 16h. This treatment was repeated with fresh β-agarase for 4 hours.
950 After digestion, 2 mL of 50 mM MES (mix of 0.5 M MES Hydrate and 0.5M MES Sodium
951 Salt, pH 5.7) was added very gently to the DNA solution and then DNA molecules were
952 combed and regularly stretched (2 kb/µm) on silanized coverslips provided by Montpellier
953 DNA Combing Facility. Linearized DNA molecules were fixed for 16 hours at 65 °C,
954 denatured in 1N NaOH for 20 minutes and blocked with 1 × PBS/ 1% BSA/ 0.1% Triton X-
955 100. Immuno-detection was done with antibodies diluted in 1 × PBS/ 1% BSA/ 0.1% Triton
956 X-100 and incubated at 37 °C in a humid chamber for 60 min. Each step of incubation with
957 antibodies was followed by extensive washes with 1 × PBS. Immuno-detection was done
958 with anti-ssDNA antibody (1/100 dilution, clone 16-19, Cat. No. MAB3034, Merck Milipore),
959 the mouse anti-BrdU antibody (1/20 dilution, clone B44, Cat. No. 347580, Becton Dickinson)
960 and rat anti-BrdU antibody (1/20 dilution, clone BU1/75, Cat. No. ABC117-7513, Abcys). The
961 following secondary antibodies were used: Alexa488 Goat anti Rat IgG (1/50 dilution, Cat.
962 No. A-11006, Invitrogen), Alexa 546 Goat anti Mouse IgG1 (1/50 dilution, Cat. No. A-21123,
963 Invitrogen) and Alexa 647 goat anti-mouse IgG2a (1/100 dilution, Cat. No. A-21241,
964 Invitrogen). Coverslips were mounted with 15 µL of Prolong Gold Antifade (Cat. No. P10144,
965 Thermo Fisher Scientific) and dried at room temperature for 16 hrs. Images were acquired
966 using Zeiss Z1 microscope, equipped with an ORCA-Flash 4.0 digital CMOS camera
967 (Hamamatsu), using a 40× objective, where 1 pixel corresponds to 325 bp (one microscope
968 field of view corresponds to ~450 kb). Acquisition and analysis of images were done by
969 MetaMorph version 7 software (Molecular Devices LLC) (RRID:SCR_002368).

970 **Measurements and statistical analysis of combed DNA molecules**

971 The speed of replication forks was estimated on individual forks, presented as IdU track
972 followed by a CldU track. Only intact (non-fragmented) forks were measured, which was
973 followed by DNA immuno-staining. The speed was calculated as the ratio between the CldU
974 track length (kilobases) and the CldU pulse duration (minutes). The IODs were measured as
975 the distance (in kilobases) between the centres of two adjacent progressing forks located on
976 an uninterrupted DNA fibre. Fork asymmetry was calculated as the ratio of the longer track
977 over the shorter track in progressing divergent forks. A ratio greater than one indicates
978 asymmetry. The GraphPad Prism 8.3.0 software (GraphPad Software Inc., La Jolla, CA, USA)
979 was used to generate graphs and to perform statistical analysis of DNA combing

980 measurements as described previously [41]. The statistical comparisons of different datasets
981 were assessed by the non-parametric Mann–Whitney two-tailed tests.

982 **DATA AVAILABILITY**

983 The SNS-seq data (FASTQ files) and the BED files containing the positions of the origins of the merged
984 and separated PCF and BSF samples from this study have been submitted to the European Nucleotide
985 Archive under the accession number PRJEB70708. The SNS-seq data analysis pipeline is available at
986 https://github.com/bridlin/finding_ORIs and https://zenodo.org/records/14289282.

987 **AKNOWLEDGEMENT**

988 Special thanks to Cedric Notredame for advising us to prepare SNS stranded libraries. We would like
989 to thank Julie A.J. Clement and Stephanie Le Gras for their help with preliminary analysis of SNS-seq
990 data. We acknowledge the Montpellier DNA Combing Facility for providing silanized coverslips. We
991 are grateful to MGX-Montpellier GenomiX (Montpellier, France) facility for performing sequencing. We
992 would also like to acknowledge the 'Eukaryotic Pathogen, Vector and Host Informatics Resource'
993 (VEuPathDB) for the integrated database, covering numerous eukaryotic pathogens [77]. We are grateful
994 to the Roscoff Bioinformatics platform ABiMS (http://abims.sb-roscoff.fr), part of the Institut Français
995 de Bioinformatique (ANR-11-INBS-0013) and BioGenouest network, for providing computing
996 resources. Thanks to Frédéric Bringaud (University Victor Segalen Bordeaux 2) for providing the *T.*
997 *brucei* 427 wild-type PCF strain and Lucy Glover (Institute Pasteur, Paris) for providing us *T. brucei* 427
998 wildtype BSF strain.

999 The work was supported by the Agence Nationale de la Recherche (ANR) within the frame of the
1000 "Investissements d'avenir" program [ANR-11-LABX-0024-01 "PARAFRAP" to YS], the Centre National
1001 de la Recherche Scientifique (CNRS), the French Ministry of Research and the Centre Hospitalier
1002 Universitaire of Montpellier. SS was recipient of a grant from the Fondation pour la Recherche
1003 Médicale (ING20160435527). MGX acknowledges financial support from France Génomique National
1004 infrastructure, funded as part of "Investissement d'Avenir" program managed by Agence Nationale
1005 pour la Recherche (contract ANR-10-INBS-09). ABiMS is founded as a part of the Institut Français de
1006 Bioinformatique (ANR-11-INBS-0013).

1007 **REFERENCES**

1008 1. Lee, C. S. K., Weiβ, M. & Hamperl, S. Where and when to start: Regulating DNA replication origin
1009 activity in eukaryotic genomes. *Nucl. Austin Tex* **14**, 2229642 (2023).
1010 2. Hulke, M. L., Massey, D. J. & Koren, A. Genomic methods for measuring DNA replication dynamics.
1011 *Chromosome Res.* **28**, 49–67 (2020).
1012 3. Leonard, A. C. & Méchali, M. DNA replication origins. *Cold Spring Harb. Perspect. Biol.* **5**, a010116
1013 (2013).
1014 4. Newlon, C. S. & Theis, J. F. The structure and function of yeast ARS elements. *Curr. Opin. Genet.*
1015 *Dev.* **3**, 752–758 (1993).
1016 5. Okuno, Y., Satoh, H., Sekiguchi, M. & Masukata, H. Clustered adenine/thymine stretches are
1017 essential for function of a fission yeast replication origin. *Mol. Cell. Biol.* **19**, 6699–6709 (1999).
1018 6. Kong, D. & DePamphilis, M. L. Site-specific DNA binding of the Schizosaccharomyces pombe
1019 origin recognition complex is determined by the Orc4 subunit. *Mol. Cell. Biol.* **21**, 8095–8103
1020 (2001).
1021 7. MacAlpine, D. M., Rodríguez, H. K. & Bell, S. P. Coordination of replication and transcription along
1022 a Drosophila chromosome. *Genes Dev.* **18**, 3094–3105 (2004).
1023 8. Balasov, M., Huijbregts, R. P. H. & Chesnokov, I. Role of the Orc6 protein in origin recognition
1024 complex-dependent DNA binding and replication in Drosophila melanogaster. *Mol. Cell. Biol.* **27**,
1025 3143–3153 (2007).

9. Stanojcic, S., Lemaitre, J.-M., Brodolin, K., Danis, E. & Mechali, M. In Xenopus egg extracts, DNA replication initiates preferentially at or near asymmetric AT sequences. *Mol. Cell. Biol.* **28**, 5265–5274 (2008).
10. Segurado, M., de Luis, A. & Antequera, F. Genome-wide distribution of DNA replication origins at A+T-rich islands in Schizosaccharomyces pombe. *EMBO Rep.* **4**, 1048–1053 (2003).
11. Mojardín, L., Vázquez, E. & Antequera, F. Specification of DNA replication origins and genomic base composition in fission yeasts. *J. Mol. Biol.* **425**, 4706–4713 (2013).
12. Eaton, M. L., Galani, K., Kang, S., Bell, S. P. & MacAlpine, D. M. Conserved nucleosome positioning defines replication origins. *Genes Dev.* **24**, 748–753 (2010).
13. Tubbs, A. *et al.* Dual Roles of Poly(dA:dT) Tracts in Replication Initiation and Fork Collapse. *Cell* **174**, 1127-1142.e19 (2018).
14. Delgado, S., Gómez, M., Bird, A. & Antequera, F. Initiation of DNA replication at CpG islands in mammalian chromosomes. *EMBO J.* **17**, 2426–2435 (1998).
15. Cadoret, J.-C. *et al.* Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 15837–15842 (2008).
16. Cayrou, C. *et al.* Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. *Genome Res.* **21**, 1438–1449 (2011).
17. Pratto, F. *et al.* Meiotic recombination mirrors patterns of germline replication in mice and humans. *Cell* **184**, 4251-4267.e20 (2021).
18. Besnard, E. *et al.* Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat. Struct. Mol. Biol.* **19**, 837–844 (2012).
19. Valton, A.-L. *et al.* G4 motifs affect origin positioning and efficiency in two vertebrate replicators. *EMBO J.* **33**, 732–746 (2014).
20. Cayrou, C. *et al.* The chromatin environment shapes DNA replication origin organization and defines origin classes. *Genome Res.* **25**, 1873–1885 (2015).
21. Comoglio, F. *et al.* High-resolution profiling of Drosophila replication start sites reveals a DNA shape and chromatin signature of metazoan origins. *Cell Rep.* **11**, 821–834 (2015).
22. Lombraña, R. *et al.* Transcriptionally Driven DNA Replication Program of the Human Parasite Leishmania major. *Cell Rep.* **16**, 1774–1786 (2016).
23. Langley, A. R., Gräf, S., Smith, J. C. & Krude, T. Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). *Nucleic Acids Res.* **44**, 10230–10247 (2016).
24. Sequeira-Mendes, J. *et al.* Differences in firing efficiency, chromatin, and transcription underlie the developmental plasticity of the Arabidopsis DNA replication origins. *Genome Res.* **29**, 784–797 (2019).
25. Akerman, I. *et al.* A predictable conserved DNA base composition signature defines human core DNA replication origins. *Nat. Commun.* **11**, 4826 (2020).
26. Prorok, P. *et al.* Involvement of G-quadruplex regions in mammalian replication origin activity. *Nat. Commun.* **10**, 3274 (2019).
27. Vitarelli, M. de O. *et al.* Integrating high-throughput analysis to create an atlas of replication origins in Trypanosoma cruzi in the context of genome structure and variability. *mBio* **15**, e0031924 (2024).
28. Lipford, J. R. & Bell, S. P. Nucleosomes positioned by ORC facilitate the initiation of DNA replication. *Mol. Cell* **7**, 21–30 (2001).
29. Berbenetz, N. M., Nislow, C. & Brown, G. W. Diversity of eukaryotic DNA replication origins revealed by genome-wide analysis of chromatin structure. *PLoS Genet.* **6**, e1001092 (2010).
30. Foss, E. J. *et al.* Identification of 1600 replication origins in S. cerevisiae. *eLife* **12**, RP88087 (2024).
31. Poulet-Benedetti, J. *et al.* Dimeric G-quadruplex motifs-induced NFRs determine strong replication origins in vertebrates. *Nat. Commun.* **14**, 4843 (2023).
32. Jansen, A. & Verstrepen, K. J. Nucleosome positioning in Saccharomyces cerevisiae. *Microbiol. Mol. Biol. Rev. MMBR* **75**, 301–320 (2011).

1078 33. Nasheuer, H. P. & Meaney, A. M. Starting DNA Synthesis: Initiation Processes during the
1079 Replication of Chromosomal DNA in Humans. *Genes* **15**, 360 (2024).
1080 34. Bielinsky, A. K. & Gerbi, S. A. Discrete start sites for DNA synthesis in the yeast ARS1 origin. *Science*
1081 **279**, 95–98 (1998).
1082 35. Foulk, M. S., Urban, J. M., Casella, C. & Gerbi, S. A. Characterizing and controlling intrinsic biases of
1083 lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-
1084 quadruplex motifs around a subset of human replication origins. *Genome Res.* **25**, 725–735 (2015).
1085 36. Castellano, C. M. *et al.* The genetic landscape of origins of replication in P. falciparum. *Nucleic*
1086 *Acids Res.* **52**, 660–676 (2024).
1087 37. Senkevich, T. G. *et al.* Mapping vaccinia virus DNA replication origins at nucleotide level by deep
1088 sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 10908–10913 (2015).
1089 38. Fenn, K. & Matthews, K. R. The cell biology of Trypanosoma brucei differentiation. *Curr. Opin.*
1090 *Microbiol.* **10**, 539–546 (2007).
1091 39. Baldauf, S. L. The Deep Roots of Eukaryotes. *Science* **300**, 1703–1706 (2003).
1092 40. Clayton, C. Regulation of gene expression in trypanosomatids: living with polycistronic
1093 transcription. *Open Biol.* **9**, 190072 (2019).
1094 41. Stanojcic, S. *et al.* Single-molecule analysis of DNA replication reveals novel features in the
1095 divergent eukaryotes Leishmania and Trypanosoma brucei versus mammalian cells. *Sci. Rep.* **6**,
1096 23142 (2016).
1097 42. Tiengwe, C. *et al.* Genome-wide analysis reveals extensive functional interaction between DNA
1098 replication initiation and transcription in the genome of Trypanosoma brucei. *Cell Rep.* **2**, 185–197
1099 (2012).
1100 43. Devlin, R. *et al.* Mapping replication dynamics in Trypanosoma brucei reveals a link with telomere
1101 transcription and antigenic variation. *eLife* **5**, e12765 (2016).
1102 44. Müller, L. S. M. *et al.* Genome organization and DNA accessibility control antigenic variation in
1103 trypanosomes. *Nature* **563**, 121–125 (2018).
1104 45. Fragkos, M., Ganier, O., Coulombe, P. & Méchali, M. DNA replication origin activation in space and
1105 time. *Nat. Rev. Mol. Cell Biol.* **16**, 360–374 (2015).
1106 46. Técher, H. *et al.* Replication dynamics: biases and robustness of DNA fiber analysis. *J. Mol. Biol.*
1107 **425**, 4845–4855 (2013).
1108 47. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator.
1109 *Genome Res.* **14**, 1188–1190 (2004).
1110 48. Valton, A.-L. & Prioleau, M.-N. G-Quadruplexes in DNA Replication: A Problem or a Necessity?
1111 *Trends Genet. TIG* **32**, 697–706 (2016).
1112 49. Marsico, G. *et al.* Whole genome experimental maps of DNA G-quadruplexes in multiple species.
1113 *Nucleic Acids Res.* **47**, 3862–3874 (2019).
1114 50. Maree, J. P., Povelones, M. L., Clark, D. J., Rudenko, G. & Patterton, H.-G. Well-positioned
1115 nucleosomes punctuate polycistronic pol II transcription units and flank silent VSG gene arrays in
1116 Trypanosoma brucei. *Epigenetics Chromatin* **10**, 14 (2017).
1117 51. Brázda, V. *et al.* G4Hunter web application: a web server for G-quadruplex prediction. *Bioinforma.*
1118 *Oxf. Engl.* **35**, 3493–3495 (2019).
1119 52. Boguslawski, S. J. *et al.* Characterization of monoclonal antibody to DNA.RNA and its application
1120 to immunodetection of hybrids. *J. Immunol. Methods* **89**, 123–130 (1986).
1121 53. Wahba, L., Costantino, L., Tan, F. J., Zimmer, A. & Koshland, D. S1-DRIP-seq identifies high
1122 expression and polyA tracts as major contributors to R-loop formation. *Genes Dev.* **30**, 1327–1338
1123 (2016).
1124 54. Briggs, E., Hamilton, G., Crouch, K., Lapsley, C. & McCulloch, R. Genome-wide mapping reveals
1125 conserved and diverged R-loop activities in the unusual genetic landscape of the African
1126 trypanosome genome. *Nucleic Acids Res.* **46**, 11789–11805 (2018).
1127 55. Godoy, P. D. de M. *et al.* Trypanosome prereplication machinery contains a single functional
1128 orc1/cdc6 protein, which is typical of archaea. *Eukaryot. Cell* **8**, 1592–1603 (2009).

56. Kim, H.-S. Genome-wide function of MCM-BP in Trypanosoma brucei DNA replication and transcription. *Nucleic Acids Res.* **47**, 634–647 (2019).
57. Zerbib, A. & Simon, I. Characterization of Unidirectional Replication Forks in the Mouse Genome. *Int. J. Mol. Sci.* **24**, 9611 (2023).
58. Bou-Nader, C., Bothra, A., Garboczi, D. N., Leppla, S. H. & Zhang, J. Structural basis of R-loop recognition by the S9.6 monoclonal antibody. *Nat. Commun.* **13**, 1641 (2022).
59. Hyrien, O. Peaks cloaked in the mist: the landscape of mammalian replication origins. *J. Cell Biol.* **208**, 147–160 (2015).
60. Tian, M. *et al.* Integrative analysis of DNA replication origins and ORC-/MCM-binding sites in human cells reveals a lack of overlap. *eLife* **12**, RP89548 (2024).
61. Marques, C. A., Dickens, N. J., Paape, D., Campbell, S. J. & McCulloch, R. Genome-wide mapping reveals single-origin chromosome replication in Leishmania, a eukaryotic microbe. *Genome Biol.* **16**, 230 (2015).
62. Marques, C. A. & McCulloch, R. Conservation and Variation in Strategies for DNA Replication of Kinetoplastid Nuclear Genomes. *Curr. Genomics* **19**, 98–109 (2018).
63. Totañes, F. I. G. *et al.* A genome-wide map of DNA replication at single-molecule resolution in the malaria parasite Plasmodium falciparum. *Nucleic Acids Res.* **51**, 2709–2724 (2023).
64. Segal, E. & Widom, J. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.* **19**, 65–71 (2009).
65. Xu, J. *et al.* Genome-wide identification and characterization of replication origins by deep sequencing. *Genome Biol.* **13**, R27 (2012).
66. Wirtz, E., Leal, S., Ochatt, C. & Cross, G. A. A tightly regulated inducible expression system for conditional gene knock-outs and dominant-negative genetics in Trypanosoma brucei. *Mol. Biochem. Parasitol.* **99**, 89–101 (1999).
67. Hirumi, H. & Hirumi, K. 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 (1989).
68. Cayrou, C., Grégoire, D., Coulombe, P., Danis, E. & Méchali, M. Genome-scale identification of active DNA replication origins. *Methods* **57**, 158–164 (2012).
69. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
70. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinforma. Oxf. Engl.* **35**, 421–432 (2019).
71. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
72. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma. Oxf. Engl.* **26**, 841–842 (2010).
73. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
74. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160-165 (2016).
75. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
76. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
77. Amos, B. *et al.* VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Res.* **50**, D898–D911 (2022).