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

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

[#] Authors contributed to the manuscript equally

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

ABSTRACT

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

Key words

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

INTRODUCTION

DNA replication is one of the most fundamental processes by which a cell creates an exact copy of its genome prior to division. In eukaryotic cells, DNA replication initiates at multiple genomic sites, which are known as origins. Despite extensive research, a comprehensive universal code that is common for all eukaryotic origins remains elusive (reviewed in [1-3]). The first isolated and characterized eukaryotic origin comes from budding yeast, *Saccharomyces cerevisiae*, where DNA replication starts from the AT-rich autonomously replicating sequence (ARS) with a loosely defined and thymidine-rich ARS consensus sequences (ACS) [4]. The other eukaryotic origins lack a unique consensus sequence. Some studies reported increased AT-content [5-11] or long poly(dA:dT) tracts as replication origins [12, 13]. Other reports described the origins as regions of a high GC-content, such as CpG islands [14-17] or regions near G-rich sequences with the potential to form G-quadruplexes (G4s) [18-26]. In addition, several epigenetic factors and chromatin environments were linked to eukaryotic origins, but without a common signature [1, 22, 27]. However, replication origins exhibit several characteristics that distinguish them from the surrounding chromatin. They appear in nucleosome-depleted regions [28-31], which are flanked on one or both sides by well-positioned nucleosomes [12, 29-32]. An active origin is characterized by the generation of two divergent single-stranded DNA molecules, each with a short RNA primer at 5′ end, known as short nascent strands (SNS) (reviewed in [33]). Size-selected SNS can be enriched and purified from DNA molecules by treatment with T4 polynucleotide kinase (T4 PNK) and λ-exonuclease, and this approach was developed for mapping of DNA replication initiation sites in yeast [34]. Thereafter, this technique has been successfully applied in different species, using approaches such as SNS-on-chip [15, 16] or by high throughput sequencing of SNS (SNS-seq) [18, 20, 22, 25, 26, 35, 36]. A recent development has improved the conventional SNS-seq method by enabling the preservation of SNS molecule directionality following sequencing [17, 37].

Trypanosoma brucei is a unicellular parasite of medical and veterinary importance that provokes deadly vector-borne diseases: sleeping sickness in humans, and nagana disease in livestock. This parasite circulates between mammalian and insect hosts, developing various life cycle stages: bloodstream form (BSF) in the mammals and the procyclic form (PCF) in the midgut of the tsetse fly [38]. The order *Trypanosomatida* diverged independently of the *Opisthokonta*, the group comprising animals and fungi [39], and has attracted attention as an original model for biological and comparative studies. These single-celled eukaryotes have some unusual genomic and epigenetic properties. Transcription is polycistronic, meaning that multiple functionally unrelated genes are transcribed into one long precursor mRNA [40]. A better understanding of the fundamental processes in these divergent eukaryotes could provide insight into evolutionary steps and pave the way for innovative medical treatments for sleeping sickness and nagana disease.

The replication of DNA in these parasites has recently become a subject of interest, [41, 42, 43]. To map and identify active DNA replication origins (hereafter referred to as "origins") in *T. brucei*, we isolated RNA-primed SNS molecules and sequenced them using a stranded protocol that

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

RESULTS

Origin mapping by stranded SNS-seq.

Stranded SNS-seq method was employed to map the origins in two distinct cell types of *T. brucei*, representing two life cycle stages of this parasite: PCF and BSF cells. The experimental workflow for purifying and enriching SNS, together with the corresponding control, is illustrated in Figure 1A. SNS-enriched samples were prepared through a two-stage process. SNS molecules were enriched first, by size-selection (0.5-2.5 kb), and second, by removing broken DNA molecules by λ-exonuclease treatments. SNS-depleted controls, prepared in parallel with SNS-enriched samples, underwent an additional RNase treatment to remove RNA primers before λ-exonuclease treatment (Methods and Figure 1A). These contained molecules resistant to RNase and/or λ-exonuclease and served as background controls. Then, stranded libraries were prepared for high-throughput sequencing (Figure 1B and Methods). Paired-end sequencing of an SNS-enriched samples and an SNS-depleted controls was performed on three biological replicates of PCF and BSF cells. Read counts after sequencing and each analysis step are shown for each sample in Supplementary Table 1 and Figure S1A.

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

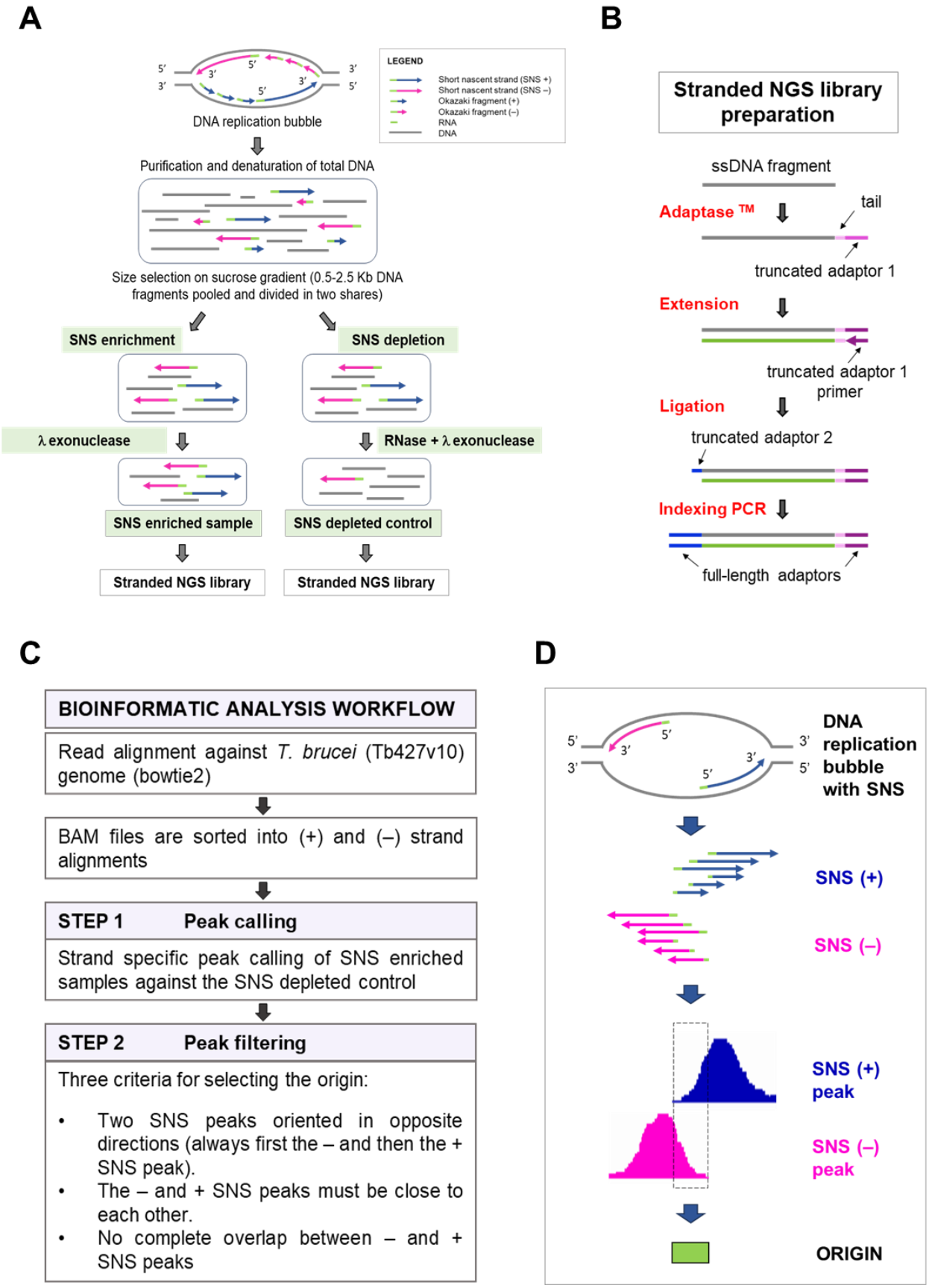


Figure 1. Experimental protocol and bioinformatics workflow for selecting the origins.

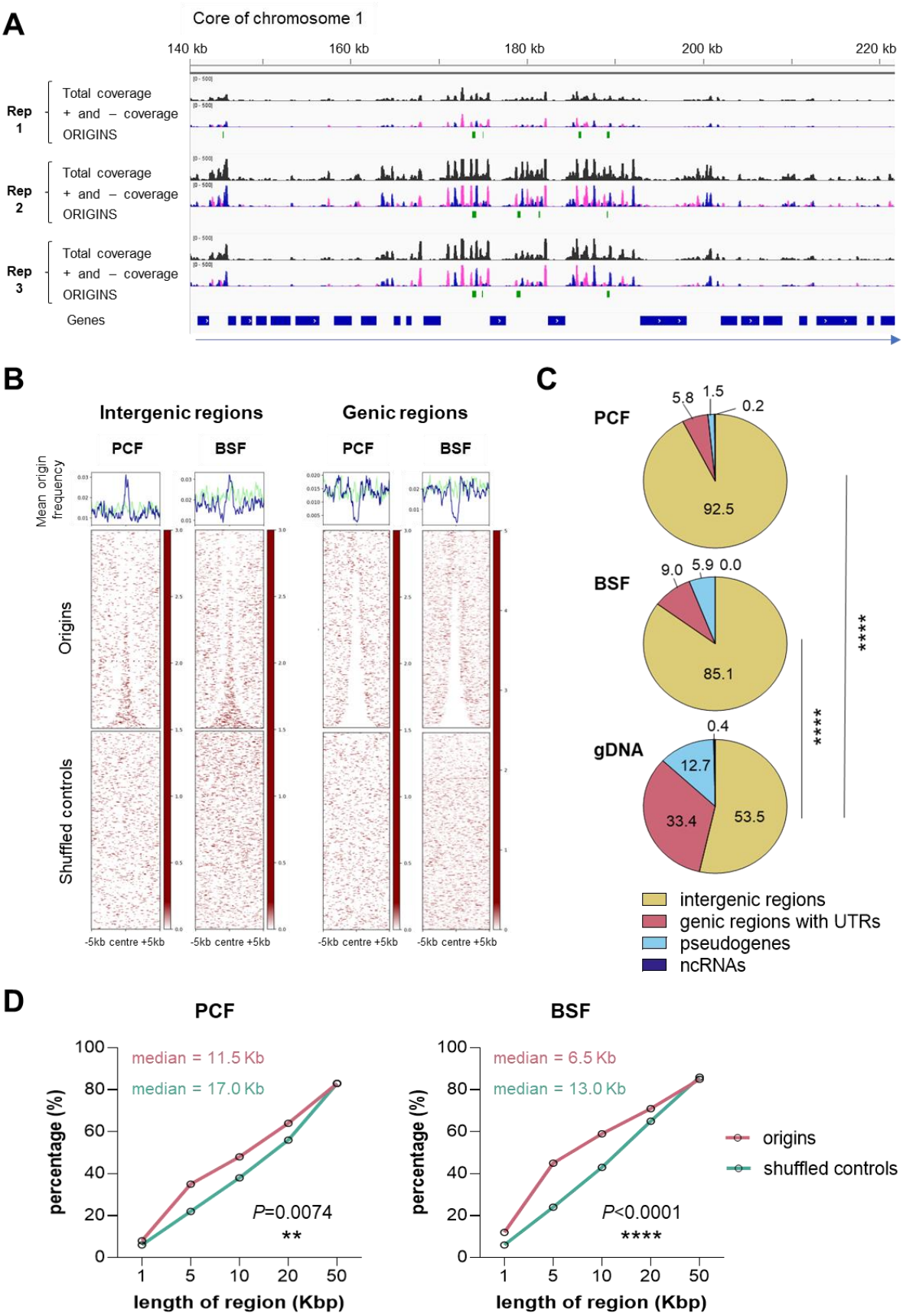
A) Schematic representation of the experimental workflow for short nascent strand (SNS) purification. A replication bubble with DNA replication intermediates and their orientation is shown at the top. The methodology employed for the generation of the SNS-enriched samples and corresponding SNS-depleted 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.

A representative screenshot of the distribution of plus and minus peaks and the positions of the mapped origins for three replicates of BSF cells is shown in Figure 2A. We found a variable number of origins in the replicates (Supplementary Figure 1D), and the overlap of origins between replicates varied from 8-64%. The distribution of the lengths of the mapped origins was estimated for each replicate (Supplementary Figure 2A), and the average length of origins was found to be approximately 150 bp. Following the mapping, the origins from the three replicates were merged, resulting in the identification of a total of 914 origins in BSF cells and 549 in PCF cells. To identify preferred genomic sites of origin activation, we analysed origin distribution around centred intergenic and genic regions. Origins were unevenly distributed, showing higher density in intergenic and lower in genic regions (Figure 2B). This intergenic enrichment was consistent across both cell types and absent in shuffled controls, which consists of randomly selected regions matched for size, number, and chromosomal distribution (Figure 2B). Subsequently, the proportion of origins in four distinct genomic regions was determined, revealing that over 92% of PCF and 85% of BSF origins map within intergenic regions (Figure 2C). This distribution was significantly different from the distribution that would be expected if the origins were distributed evenly across the genome (*P*-value$<$0.0001) (Figure 2C).

We then wanted to see if origins were organised as solitary sites or grouped together, so we examined how they clustered within genomic regions of increasing length. Our results showed that *T. brucei* origins were predominantly organised in clusters, with 85.8% of PCF origins and 83.2% of BSF origins organized in clusters of 50 kbp (Supplementary Figure 2B). Empirical cumulative distribution function (ECDF) was calculated for both cell types, and a statistically significant difference was confirmed between the origins and shuffled controls clustering (Figure 2D). The results showed that 50% of the PCF origins were clustered within 11.5 kb regions (shuffled control median was 17 kb), while 50% of the BSF origins were clustered within 6.5 kb regions (shuffled control median was 13 kb) (Figure 2D). This means that origins tend to be activated within initiation zones or clusters rather than as solitary initiation sites in *T. brucei*. This finding is consistent with previous research in other organisms, which has shown that origins are unevenly distributed and tend to form clusters of different lengths [20, 22, 25]. Next,

147 the intersection of the two sets of origins revealed that 45% of origins were common between
148 the PCF and BSF cell types (Supplementary Figure 2C). The number of origins in both cells was
149 compared using normalised mapped reads (Methods), and after merging the reads from the
150 three replicates and normalization, we showed that BSF cells contained almost twice as many
151 origins as PCF cells (merged BSF: 844 origins and averaged normalised subsamples of PCF: 454
152 origins) (Supplementary Figure 2D).

153 We then calculated the inter-origin distances (IODs) distribution at the population level for
154 both cell types. The calculated population-based IODs were shorter in BSF (median 15.3 kb)
155 compared to PCF cells (median 21.4 kb) and this difference was statistically significant
156 (Supplementary Figure 2E). Very long IODs were rarely observed, with maximum distances of
157 354 and 488 kilobases between PCF and BSF origins, respectively (Supplementary Figure 2E).



158

159 **Figure 2. Genomic distribution of mapped origins.**

A) A representative screenshot from Integrative Genomics Viewer (IGV) presenting the read coverage and the positions of the mapped origins in a ~80 kb long genomic region of the core chromosome 1. The three replicates of BSF cells are presented. The total read coverage is in black, the minus and plus strand read coverage are in pink and blue, respectively. The green bars represent the mapped origins. The blue bars show the position of the genes and the direction of transcription of the polycistronic unit is indicated by the blue arrow below. B) Heatmaps present the distribution of the mapped origins and shuffled controls within centred genic and intergenic regions (± 5 kb). Shuffled controls present random genomic regions chosen with respect to the size and chromosomal distribution of origins. C) The pie charts show the proportions of origins within four indicated genomic regions. The gDNA presents the proportions of four genomic regions within genomic sequence. The *P*-values were computed using Chi-square tests, which involved a comparison of the absolute numbers of indicated categories for each pair of datasets (**** - $P<0.0001$). D) Empirical cumulative distribution function (ECDF) showing the proportion of pairwise distances between origins (pink) or between shuffled control regions (green). The *P*-values were computed using Chi-square tests.

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

Stranded SNS-seq and DNA molecular combing are two complementary techniques. Stranded SNS-seq provides a genome-wide overview of active origins within a given cell population. However, it should be noted that the usage of these origins may vary between individual cells within the same population (reviewed in [44]). DNA combing provides an overview of DNA replication parameters at the single-molecule level. In this study, we used DNA combing to compare the replication of PCF and BSF cells. Replicating cells were labelled with two modified nucleosides, followed by immuno-detection (Methods). After immuno-detection, the three discrete signals were discerned on combed DNA. The red signal, followed by green, indicated the direction of replication forks, while the blue signal represented the unmodified DNA (Figure 3A). Origins were positioned at the centre of two divergently oriented replication forks. DNA replication parameters, including replication fork speed, IODs, and fork asymmetry, were determined as previously described [41]. Measurements from three biological replicates were pooled and column statistics are shown in Supplementary Table 2. The median replication fork velocity was 1.73 and 2.44 kb/min for PCF and BSF, respectively, while the median IOD was 152 and 213 kb for PCF and BSF, respectively (Supplementary Table 2). The median speed of replication forks was 1.4 times higher, and the median IOD was 1.4 times longer in BSF compared to PCF cells, with these differences between the two cell types being statistically significant (Figure 3B and Figure 3C). Moreover, we compared the frequency of asymmetric replication forks, defined as bidirectional forks that progress from the origin at unequal rates, between the PCF and BSF cell types and found no statistically significant differences (Figure 3D). Given the high sensitivity of DNA combing measurements to the DNA fibre lengths [45], a comparison of the analysed DNA fibre lengths was performed as a control. This analysis showed that there was no statistically significant difference in the length of the analysed DNA molecules between the two cell types (Figure 3E, Supplementary Table 2). This control confirmed that the observed differences in velocity and IODs were not an artefact of the different lengths of the analysed DNA molecules but rather reflected genuine differences between the replication of PCF and BSF cells.

202 The approximate number of active origins in a single cell can be estimated by dividing the
203 length of genomic sequences (~50 Mb) by the mean or median IODs obtained by DNA
204 combing. It was thus estimated that the range of active origins in a single cell was 279-329 for
205 PCF and 228-235 for BSF cells (Supplementary Table 2). The stranded SNS-seq method
206 revealed 549 origins for PCF and 914 origins for BSF under stringent origin selection, which
207 represent a fraction of all origins. Despite the inability of either method to determine the exact
208 number of origins, the number of origins in a single cell tends to be smaller than in a
209 population of cells. This phenomenon, known as flexible origin usage, is a well-documented
210 feature of eukaryotic origins (reviewed in [44]).

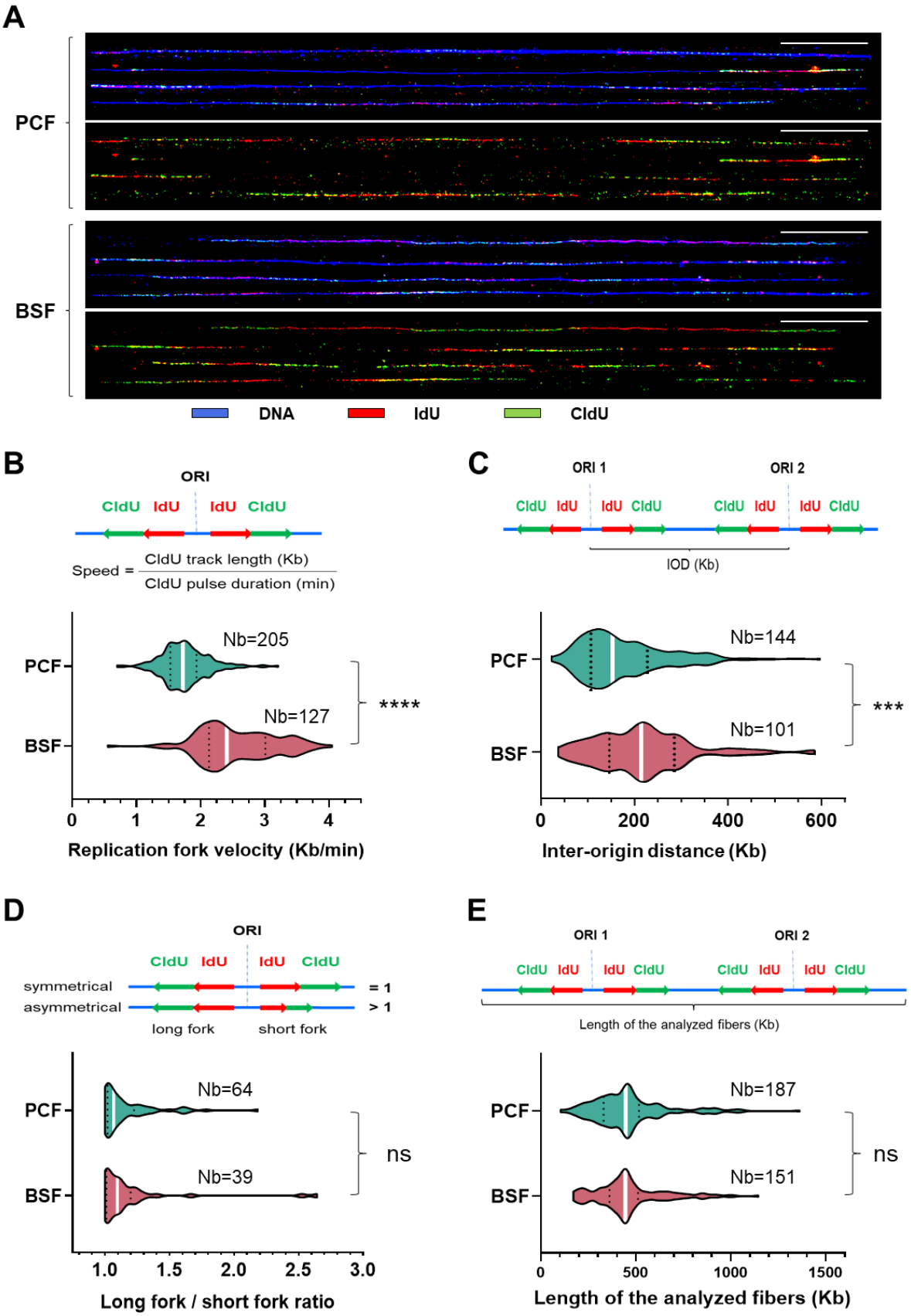


Figure 3. Comparison of DNA replication parameters between PCF and BSF cells by DNA molecular combing.

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 consensus motif was identified (data not

259 shown). The sequences surrounding the origins were also analysed using the WebLogo
260 software [46], and a graphical representation of nucleotide occurrence unambiguously confirmed
261 an unequal nucleotide distribution, with an enrichment of A and G upstream and T and C
262 downstream of the origins (Supplementary Figure 3B). We then looked at the distribution of
263 polynucleotides of different lengths around the origins. Poly(dA)-rich sequences were enriched
264 upstream and poly(dT)-rich sequences downstream of the origins, a pattern absent in shuffled
265 controls (Figure 4B, Supplementary Figure 3C). Interestingly, the upstream region showed an
266 enrichment of up to four consecutive Gs, while the downstream region showed an enrichment
267 of up to four consecutive Cs (Figure 4B). We also checked nucleotide and polynucleotide
268 distributions around PCF and BSF origins separately and found that the distributions were
269 similar in both cell types and equivalent to the merged set of origins (Supplementary Figure
270 4A and Supplementary Figure 4B). Accordingly, an origin can be defined as a genomic site with
271 a specific uneven distribution of nucleotides, characterised by an overrepresentation of
272 poly(dA)s and poly(dG)s upstream, and poly(dT)s and poly(dC)s downstream of the origin
273 centre.

274 Subsequently, the proportions of origins that lacked or possessed four As upstream and/or
275 four Ts downstream of the centre were calculated and compared with the intergenic shuffled
276 controls (Figure 4C, left graph). The same calculation was performed for eight As and Ts (Figure
277 4C, right graph). The findings revealed that 1149 origins exhibited the presence of four As
278 upstream and four Ts downstream of the centre, representing a proportion of ~94% compared
279 to ~78% for the intergenic shuffled controls. Similarly, 450 origins (~37%) possessed eight As
280 upstream and eight Ts downstream of the centre, compared to only 114 (~9%) intergenic
281 shuffled controls with this pattern. Accordingly, the distribution of poly(dA) and poly(dT)
282 observed for origins was found to be significantly different from that observed for the
283 intergenic shuffled controls (Figure 4C).

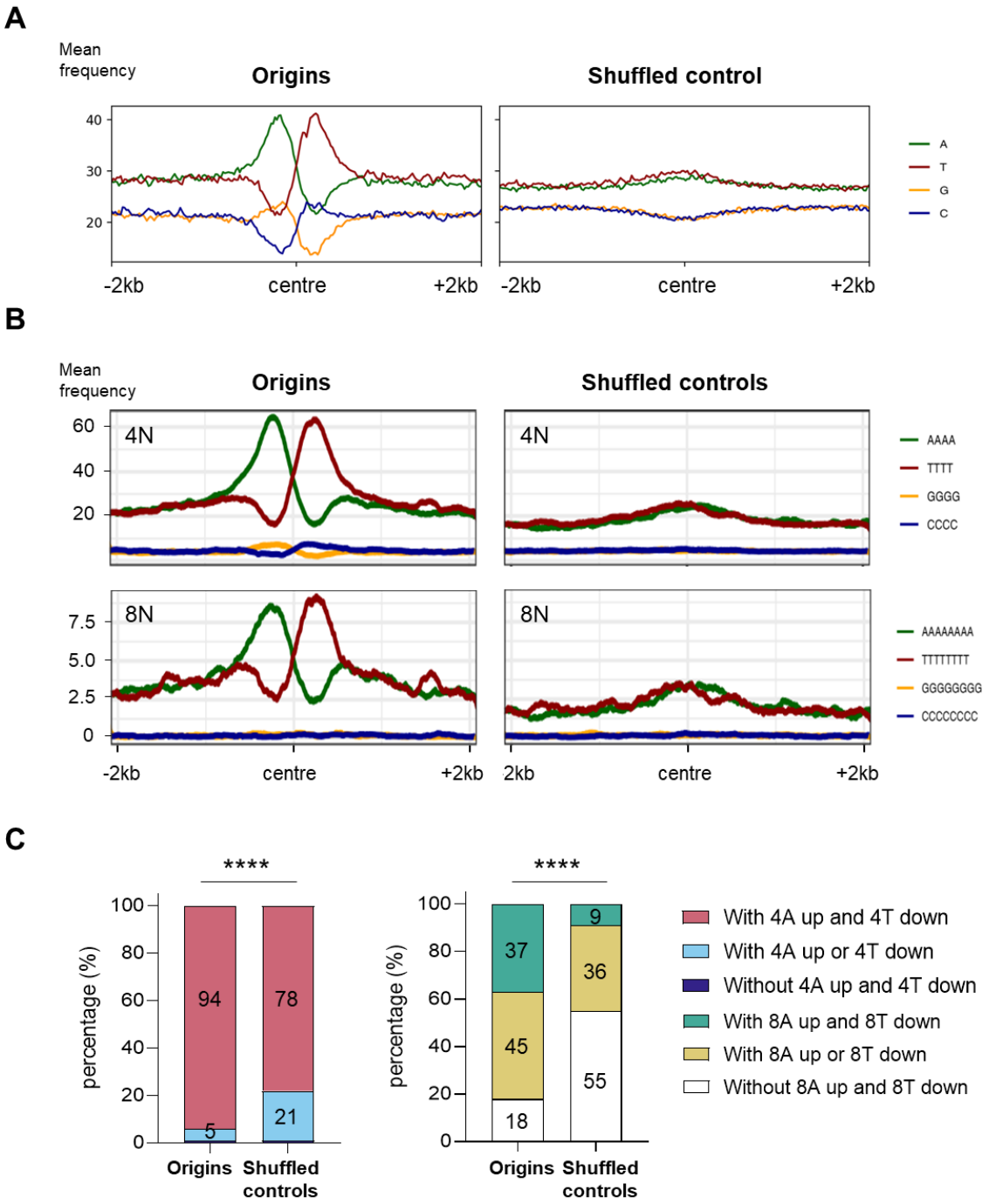


Figure 4. Spatial organisation of nucleotides and polynucleotides in the vicinity of origins.

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

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

296 G4s, DNA secondary structures with non-canonical Watson-Crick base pairing, have recently
297 emerged as potential regulators of genome replication [47]. Given the previously reported
298 presence of G4 structures in the vicinity of some eukaryotic origins [18, 19, 21, 26], we investigated
299 the correlation between G4s and origins of *T. brucei*. To do this, we used two sets of G4s that
300 had previously been determined by G-quadruplex sequencing (G4-seq) [48]. One set of G4s was
301 obtained under physiological conditions, while the other was obtained in the presence of
302 pyridostatin (PDS), a drug that binds and stabilises G4s [48]. The SNS-seq origin dataset was
303 intersected with stranded G4-seq datasets obtained in both conditions (Methods), resulting in
304 a notable enrichment of G4s in the vicinity of the origins (Figure 5A and Supplementary Figure
305 5A). A distinctive pattern of G4 distribution was observed, whereby the G4s on the plus strand
306 were found to be enriched upstream, while the G4s on the minus strand exhibited an
307 enrichment downstream of the centres of origin (Figure 5A and Supplementary Figure 5A).
308 This pattern of G4 enrichment was absent in the intergenic shuffled controls (Figure 5A and
309 Supplementary Figure 5A). The occurrence of G4s near the origins varied, with a discernible
310 increase of G4s near the centre observed under PDS-stabilised conditions in comparison to
311 physiological conditions (Figure 5A and Supplementary Figure 5A).

312 Our subsequent objective was to determine the proportion of origins that possessed G4s on
313 one or both sides and those that lacked G4s. To achieve this, the region of ±2 kb around the
314 centres of origins was analysed and compared to shuffled controls. The result showed that the
315 origins displayed significantly different G4 distributions compared to the shuffled controls
316 ($P<0.0001$) (Figure 5B). Furthermore, the proportions were significantly different in PDS-
317 stabilised compared to physiological conditions (Figure 5B). A set of origins with G4s on both
318 sides constituted a minor population under physiological conditions (9%), whereas this was a
319 major population in PDS-stabilised conditions (39%). The proportion of origins with G4 on one
320 or both sides increased significantly in PDS-stabilised condition (74%), compared to the
321 physiological condition (47%) (Figure 5B). Analyses of the presence of G4s in the region ±0.5
322 kb around the centres of origins yielded comparable results (Supplementary Figure 5B).
323 Quantification of G4 distribution around the PCF and BSF origins, was similar in the two cell
324 types, and comparable to the merged set of origins (Supplementary Figure 5C).

325 To gain a deeper insight into the distribution of origins and the G4 structures, an empirical
326 cumulative distribution function (ECDF) was computed. The ECDF demonstrated that the
327 distances between the origins and the closest G4s were smaller than those observed for the
328 shuffled controls (for physiological G4s condition - median of 2010 bp for origins versus
329 median of 4154 bp for intergenic shuffled controls and for stabilised G4s condition – median
330 of 329 bp for origins versus median of 1223 bp for intergenic shuffled controls) (Figure 5C).
331 Half of the origins had a stabilised G4 structure within ±329 bp of the centre, while 80% of the
332 origins had the same structure within ±2575 bp of the centre. Therefore, a strong association
333 was identified between origins and physiological G4 structures (odds ratio = 1.87; *P*-value =
334 0), as well as for stabilised G4s structures (odds ratio = 2.13; *P*-value = 0).

335 Our results demonstrated that poly(dA) enriched sequences and G4s were enriched upstream
336 of origins on the plus strand, and downstream on the minus strand (Figure 4B and Figure 5A).
337 Therefore, we wanted to determine the spatial relationship between these two elements. To
338 achieve this, the strand separated poly(dA) enriched sequences and the strand separated
339 experimental G4s [48] were intersected at centred origins and shuffled controls (Methods). The
340 results demonstrated that the peaks of the experimental G4s and poly(dA) sequences
341 overlapped near the centres of origins in a strand specific manner that was notably absent in
342 the intergenic shuffled controls (Figure 5D). Our findings revealed a previously unidentified
343 element of origin: the enrichment of poly(dA) and G4 structures exhibited an important
344 association upstream of the centre on the plus strand and downstream of origin on the minus
345 strand (Figure 5D). It was observed that the poly(dA) peaks exhibited summits between ±100
346 to 300 nt, while the G4 peaks exhibited summits in the range of ±100 to 200 nt from the
347 centres of origins (Figure 5D). Such regions were designated as poly(dA)/G4-enriched
348 elements of origins.

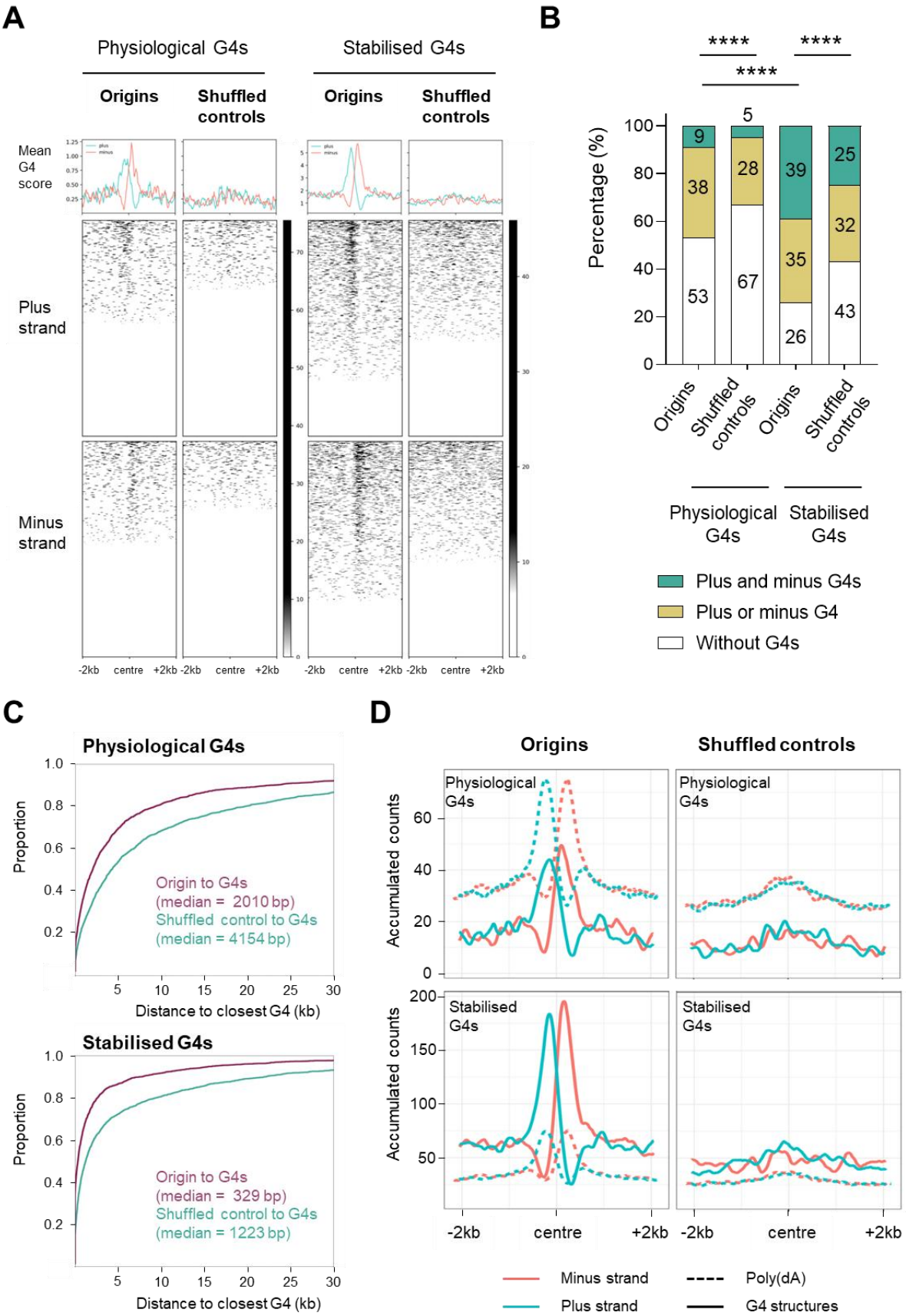


Figure 5. Distribution of G4 structures in the vicinity of origins.

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 [48] were overlapped with origins mapped by stranded SNS-seq. Mean G4 score presents average G4 score [48] per 20 bp window. B) The proportions of origins that lack or possess G4 structures on one or both sides of the centre were determined. The ±2 kb window from the centre was analysed for the presence of G4s. Two sets of experimental G4 structures [48] were subjected to analysis in comparison with intergenic shuffled controls. The P-values were calculated using the Chi-square test, which involved a comparison of the absolute numbers of indicated categories for each pair of datasets (**** - P<0.0001). C) Empirical Cumulative Distribution Function (ECDF) of the distances between the origins and the closest physiological and stabilized G4 structures [48]. Median distances are indicated. D) The profile plots illustrate the distribution of the poly(dA) sequence (AAAA) (dashed lines) and the experimental G4s (physiological and PDS drug stabilized) [48] (solid lines) around centred origins and intergenic shuffled controls (±2 kb). The analysis was conducted in the Tbb TREU927 reference genome (Methods). The plus strand is represented in blue and the minus strand in pink. A smoothing function was employed to calculate the accumulated counts of the AAAA and G4s per position within a 100 bp window (50 bp up- and downstream from the position).

The origins exhibited varying nucleosome occupancy.

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

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

395 significantly more LNO flanked by two HNO, at the expense of the no HNO with LNO and the
396 no HNO no LNO categories.

397 Further investigation was undertaken to determine the spatial relationship between
398 poly(dA)/G4 enriched sequences and nucleosome distribution. To achieve this, strand
399 separated poly(dA) enriched sequences and predicted G4s were overlapped with the MNase-
400 seq data [49] (Methods). The plot obtained after the intersections revealed that the LNO region
401 was positioned between the plus and minus poly(dA)/G4 summits (Figure 6D). Additionally,
402 the HNO peaks were found to exhibit overlap with the poly(dA) and G4 peaks and to surround
403 the origins (Figure 6D). Furthermore, these intersections permitted the estimation of the
404 averaged distance of the above-mentioned elements from the centre of origins. The LNO
405 region extended to approximately ±100 to 150 nt from the centre, thereby defining an area of
406 approximately 200 to 300 bp in the centre of origin. The two HNO regions were located
407 between 150 and 500 nt upstream and downstream from the centre of origin (Figure 6D).

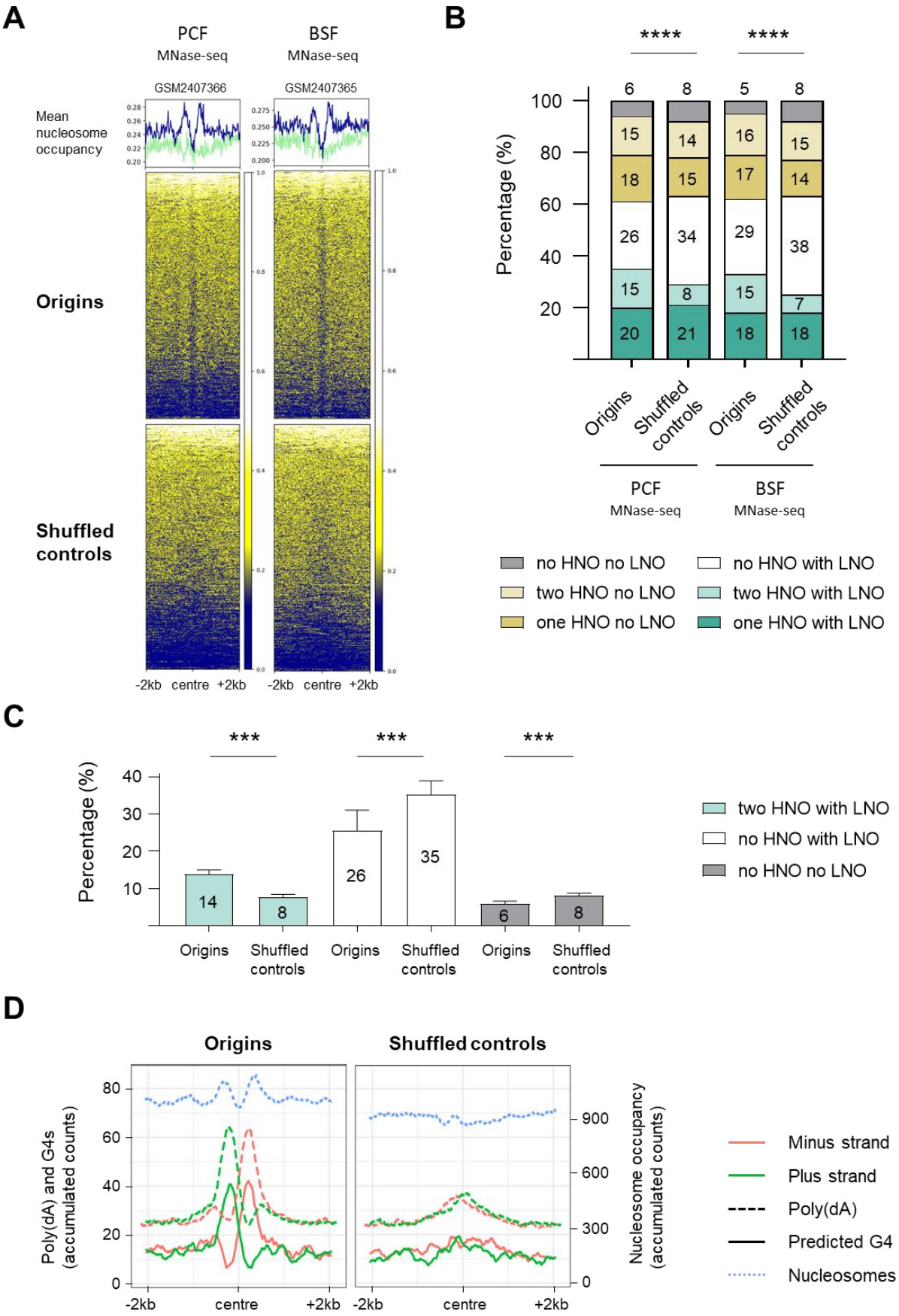


Figure 6. The distribution of nucleosomes around the origins in *T. brucei*.

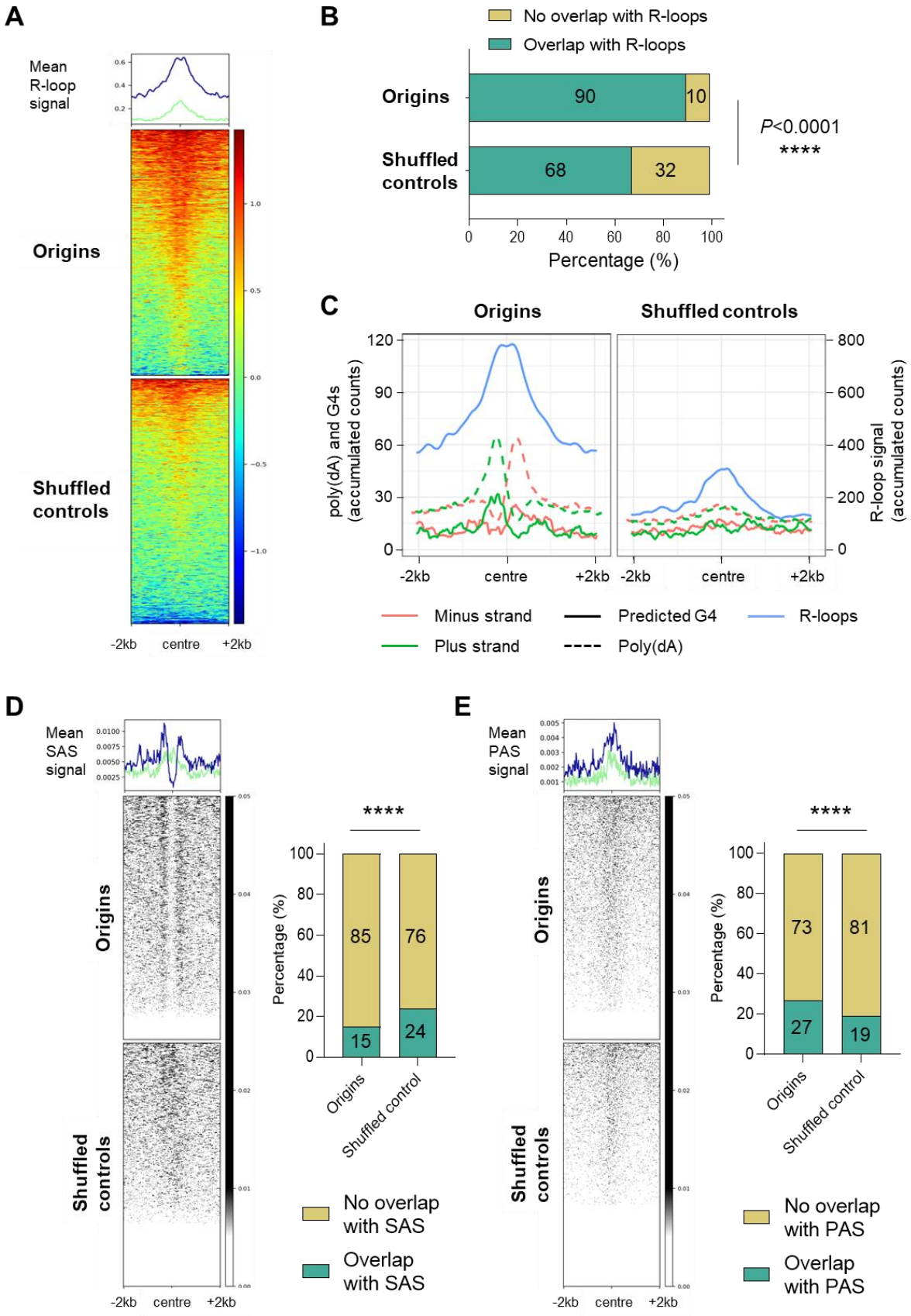
A) The profile plot and heatmaps show the distribution of nucleosomes (MNase-seq data) detected in PCF and BSF cells [49] around the centred origins (blue line) and the intergenic shuffled controls (green line) (± 2 kb) in Tb Lister 427 reference genome. One replicate of MNase-seq data [49] for PCF (GSM2407366) and one replicate for BSF cells (GSM2407365) is shown here, the other replicates are presented in Supplementary Figure 6A. Mean nucleosome occupancy presents average nucleosome score (dyad value) [49] per 20 bp window. B) The percentage of origins and intergenic shuffled controls with or without high nucleosome occupancy (HNO) and low nucleosome occupancy (LNO) regions in the indicated combinations. Quantification was performed for the same replicates as indicated in the Figure 6A. The *P*-values were calculated using the Chi-square test, which involved a comparison of the absolute numbers of indicated categories for each pair of datasets (**** - $P<0.0001$). C) Percentage of the three specified categories of HNO and LNO combinations that were significantly different between origins and shuffled controls. The remaining three categories were not statistically significant. The numbers indicate mean percentage values. The bars indicate standard deviations. Mann Whitney two-tailed test was performed to calculate *P*-values (*** - $P<0.001$). D) The profile plots show the distribution of the poly(dA) sequence (AAAA) (dashed line), the predicted G4 structures (solid line) and the nucleosome distribution of the replicate GSM2407365 from BSF cells [49] (blue dotted line) around centred origins and intergenic shuffled controls (± 2 kb). The G4s were predicted by the G4Hunter application [50] in the Tb Lister 427 reference genome. The plus strand is represented in green, and the minus strand in pink. The values on the plots present the accumulated counts of the AAAA, G4s and nucleosome occupancy scores per position within a 100 bp window (50 bp up- and downstream from the position).

The origins colocalized with intergenic R-loops enrichment sites.

R-loops are defined as three-stranded structures consisting of RNA:DNA hybrids and a displaced single-stranded DNA. It has been demonstrated that RNA:DNA hybrids occur naturally during cellular functions such as transcription and replication [51]. DRIP-seq is a widely used technique for genome-wide mapping of R-loops. This method is based on the immuno-precipitation of fragmented genomic DNA by the S9.6 monoclonal antibody, which specifically recognises RNA:DNA hybrids [52, 53]. A recent study employed DRIP-seq to map R-loops in *T. brucei* BSF cells, revealing their enrichment within intergenic regions, between coding sequences of polycistronic transcription units, and coinciding with sites of polyadenylation and nucleosome depletion [54]. Considering the established role of short RNA primers in initiating DNA synthesis at the leading and lagging strands, an examination of the colocalization of the origins and previously published DRIP-seq data [54] was conducted (Methods). A significant accumulation of R-loops or RNA:DNA hybrids was identified around the origins, with two distinct summits located close to the origin centre (Figure 7A). An R-loop enrichment was also evident in the vicinity of the centre of the intergenic shuffled controls (Figure 7A), which is consistent with previous reports indicating a high prevalence of R-loops in intergenic regions [54]. However, this peak was observed to be lower in comparison to the peak observed near centres of origins. Furthermore, two upper summits were absent in the shuffled controls (Figure 7A). Next, we quantified, the number of origins and shuffled controls that overlapped with R-loops [54]. The results demonstrated that 90% of the origins and 68% of the shuffled controls exhibited overlap with R-loops (Figure 7B). Consequently, a statistically significant correlation was observed between the origins and R-loops (odds ratio = 4.17; *P*-value $2.87E^{-142}$), suggesting a link between RNA:DNA hybrids formation and origin activity. We also checked the distribution of R-loops detected in BSF cells around PCF and BSF origins separately and found that the distribution of RNA:DNA hybrids was similar around BSF origins and equivalent to the merged set of origins (Supplementary Figure 7A). The distribution of RNA:DNA hybrids was also enriched around the centres of PCF origins with one of the upper

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

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



466

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

469 A) The profile plots and heatmaps show the distribution of R-loops [54] around the centred origins (blue line) and the
470 intergenic shuffled controls (green line) (± 2 kb). Mean DRIP-seq signal presents average DRIP-seq signal [54] per 20
471 bp window. B) The proportion of origins and shuffled controls that overlap with R-loops (intersection without
472 window). The *P*-values were calculated using Fisher's exact (two-sided) test, which involved a comparison of the
473 numbers of indicated categories for each pair of datasets (**** - *P*<0.0001). C) The profile plots illustrate the
474 distribution of the stranded poly(dA) sequence (AAAA) (dashed line), predicted G4s (solid line) and R-loops [54] (solid
475 blue line) around centred origins and shuffled controls (± 2 kb). The plus strand is represented in green, and the
476 minus strand in pink. The 13,409 G4s were predicted by the G4Hunter application [50] in the Tb Lister 427-2018
477 reference genome. The values on the plots present the accumulated counts of the AAAA, G4s and R-loop signals
478 per position within a 100 bp window (50 bp up- and downstream from the position). D) Left panel: the profile plots
479 and heatmaps show the distribution of transcription splice acceptor sites (SAS) around the centred origins (blue
480 line) and the intergenic shuffled controls (green line) (± 2 kb). Intersection was performed in the Tbb TREU927
481 reference genome for the 11 megabase chromosomes. Right panel: the proportion of origins and shuffled controls
482 that overlap with SAS (intersection without window). The *P*-value was calculated using Fisher's exact (two-sided)
483 test, which involved a comparison of the numbers of indicated categories for each pair of datasets (**** - *P*<0.0001).
484 E) Left panel: the profile plots and heatmaps show the distribution of transcription polyadenylation sites (PAS)
485 around the centred origins (blue line) and the intergenic shuffled controls (green line) (± 2 kb). Intersection was
486 performed as in Figure 7D. Right panel: the proportion of origins and shuffled controls that overlap with PAS. The
487 *P*-values were calculated using Fisher's exact (two-sided) test, which involved a comparison of the numbers of
488 indicated categories for each pair of datasets (**** - *P*<0.0001).

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

491 We compared the replication origins identified by stranded SNS-seq with the data obtained
492 by Marker frequency analysis coupled with deep sequencing (MFA-seq) [42, 43]. MFA-seq
493 compares read coverage between replicating and non-replicating cells and identifies regions
494 with increased coverage in S phase as replication origins. In *T. brucei*, MFA-seq identified broad
495 regions of a few hundred kilobases [42, 43], whereas the origins detected by stranded SNS-seq
496 are about 150 bp long. We used six MFA-seq datasets [42, 43] (Methods) and identified the
497 previously published 47 MFA-seq peaks [43]. We then intersected the MFA-seq datasets with the
498 SNS-seq dataset (Supplementary Figure 8A) to assess the extent of the overlap
499 (Supplementary Figure 8B). We found 28-42% stranded SNS-seq origins overlapped with early
500 and 43-55% overlapped with late S-phase MFA-seq replicated regions (Supplementary Figure
501 8B).

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

512 **Spatial coordination between the activity of the origin and transcription.**

513 DNA replication in trypanosomatids occurs in a particularly challenging environment, as most
514 of their genomes are constitutively transcribed. Arrays of genes oriented in the same direction
515 are transcribed by RNA polymerase II into polycistronic transcription units (PTUs), which can
516 contain hundreds of genes (reviewed in [40]). The primary transcript is processed into individual
517 mRNAs by trans-splicing of a capped 'spliced leader' at the splice acceptor site (SAS) and by
518 polyadenylation at the polyadenylation sites (PAS) (reviewed in [40]). To gain insight into the
519 relationship between origins and SAS and PAS, we intersected our dataset with the unified SAS
520 and PAS datasets for the core chromosomes (Methods). The intersection of origins with SAS
521 showed that the centre of origins was depleted of SAS, while some origins had SAS flanking
522 the centre and the SAS signal was enriched ±200-500 nt from the centre of origins (Figure 7D,
523 left panel). Only 15% of the origins overlapped with SAS compared to 24% of the shuffled
524 controls, and this difference was statistically significant (Figure 7D, right panel). The
525 intersection of the origins with PAS revealed an accumulation of PAS near the origin centres.
526 A similar, though less pronounced, accumulation was also observed in the shuffled controls
527 (Figure 7E, left panel). Quantification showed that 27% of origins overlapped with PAS
528 compared to 19% of shuffled controls and this difference was statistically significant (Figure
529 7E, right panel).

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

538 **DISCUSSION**

539 We used stranded SNS-seq to identify active origins in the unicellular eukaryote, *T. brucei*. Our
540 methodology is comparable to previously published origin-derived single-stranded DNA
541 sequencing (Ori-SSDS) [17] but includes few differences. The stranded library preparation
542 process was different in these two approaches, and the SNS-depleted control was included in
543 the origin mapping process as a background control. The key benefit of stranded over
544 conventional SNS-seq is the ability to identify origins with greater specificity by eliminating
545 peaks that do not align with the proper orientation of SNS of an active origin. In addition, this
546 method maps the origins between two divergent SNS peaks, allowing the determination of
547 the centre of origins or replication initiation sites, which was essential for the discovery of the
548 positions and orientations of the elements involved in the topology of origins. Origins mapped
549 between two divergent SNS peaks showed the origin-specific nucleotide distribution around
550 the centre (Figure 4A, Supplementary Figure 9A). On the other hand, if we apply conventional
551 SNS-seq origin mapping to our data and map origins in the middle of peaks without knowing
552 their orientation, the origin-specific nucleotide distribution is lost (Supplementary Figure 9B).

553 Therefore, the primary benefits of our approach include the ability to centre origins between
554 two divergent SNS peaks and to eliminate false positive peaks.

555 Using this stringent origin selection procedure, we have selected and analysed a subset of all
556 origins. Some weak or inefficient origins may be missed due to low activity, resulting in low
557 read coverage. It is therefore very likely that the number of origins of this parasite is higher
558 than the number reported here. We also examined the origins separately in the two life stages
559 demonstrating that the origin structures in the PCF and BSF cells were equivalent and justifying
560 the merging of these two sets of origins into one. Under this stringent origin mapping
561 condition, we found that 45% of PCF origins were shared with BSF cells (Supplementary Figure
562 2C). However, we decided not to compare the PCF and BSF sets of origins in more detail
563 because comparing non-exhaustively mapped origins in these two life cycle stages could be
564 misleading.

565 Our analysis revealed no consensus sequence of origins, which is consistent with the findings
566 of most other studies of eukaryotic origins, including trypanosomatids [22, 27, 42, 43]. Nevertheless,
567 the genomic sites that function as origins are not random sequences. Thanks to the stranded
568 SNS-seq approach, we were able to determine the positions and strand-specificity of some
569 genetic elements and to present a comprehensive model of an origin (Figure 8). The results of
570 our study showed that DNA replication starts at genomic sites where adenine and thymine are
571 overrepresented and arranged in a specific pattern. On the plus strand, poly(dA) enriched
572 sequences were situated upstream, while poly(dT) enriched sequences were located
573 downstream of the origin's centre (Figure 4B and Figure 8). The G4s near the centre of origins
574 were embedded in poly(dA) rich sequences and the summits of G4 peaks covered 100-200 nt
575 up and downstream from the centres (Figure 5D, Figure 8). The centres of origin were identified
576 as LNO regions, which had an average length of 200–300 nt and were flanked by HNO regions
577 located from ±150 to ±500 bp from the centre (Figure 6D, Figure 8). RNA:DNA hybrid
578 enrichment revealed two distinct summits, one upstream and one downstream of the centre,
579 with a range spanning ±200 nucleotides (Figure 7A and Figure 8). Our model suggests that
580 the leading strands of DNA replication, or SNS molecules, are synthesised using the
581 poly(dA)/G4-rich strands as templates.

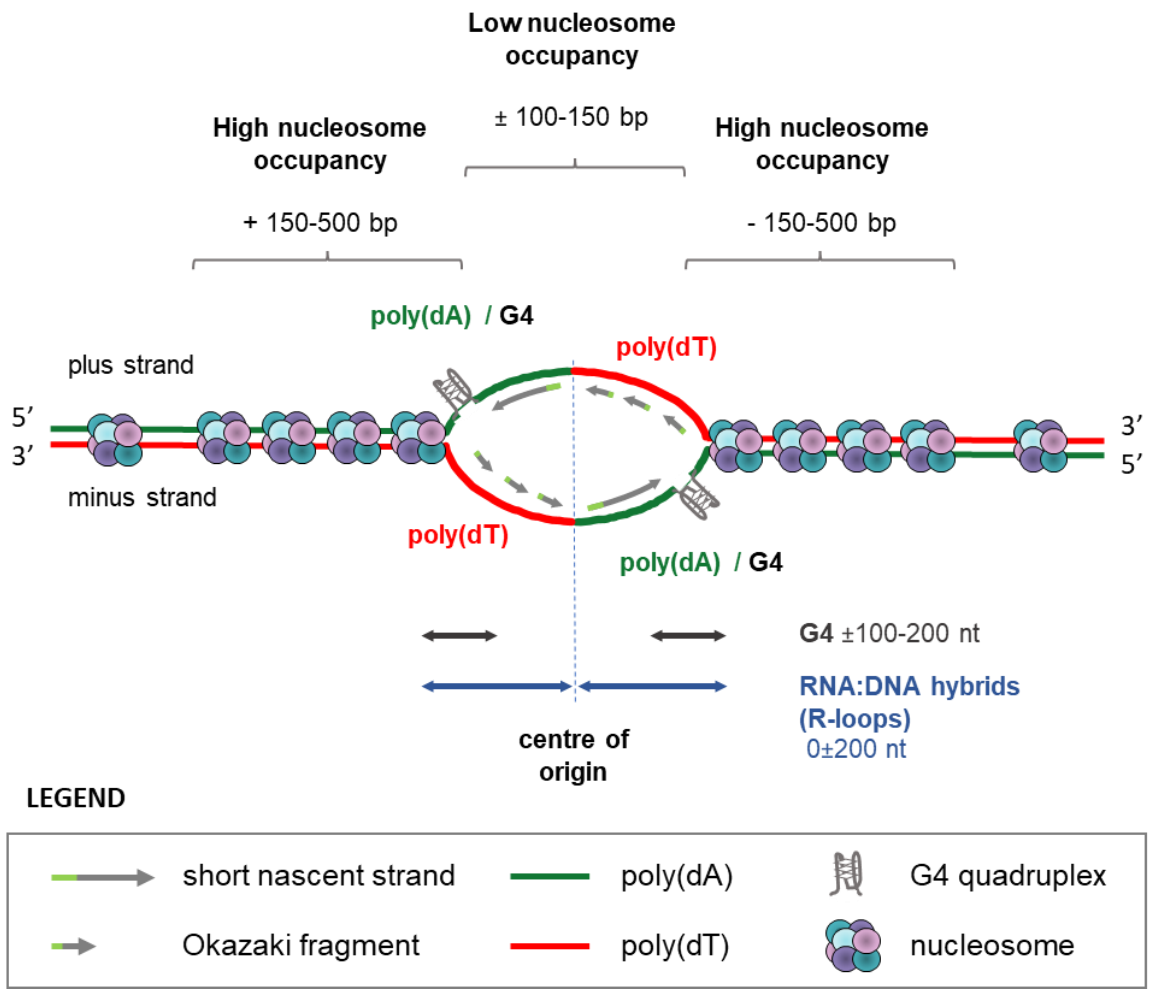


Figure 8. A proposed model of origin with the position of different genetic elements and nucleosome occupancy.

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

Our analysis indicated that G-rich sequences near origins tend to form G4 structures (Figure 5A-C, Supplementary Figure 5B-C). The detection of more G4s under PDS-stabilized conditions compared to physiological ones suggests that there is an important potential for G4s formation near origins. G4 formation seems to be very transient and therefore not easy to detect unless stabilized by drug. It seems that G4 structures near origins undergo rapid changes, either forming or dissolving in response to the replication dynamics. We therefore propose that the formation of G4s plays a regulatory role in the activation of replication origins. Some of the recent publications suggested that G4s may serve as replication fork barriers and play a role in transient replication fork stalling [21, 56]. The enrichment of G4

602 structures on strands acting as templates for leading strand synthesis implies that the
603 formation of G4 structures could potentially pose or arrest replication fork progression.

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

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

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

642 suggested by MFA-seq at the population level (~600 kbp) [42]. Thirdly, as MFA-seq calculates
643 the ratio of read coverage between the S and G2 phases, it may lack the sensitivity required to
644 detect origins used by fewer than 25% of cells in the population [60]. Similarly, stringent origin
645 detection using stranded SNS-seq identified a subset of origins. Therefore, comparing two
646 incomplete datasets has its limitations and may not be relevant.

647 Validation of origins is generally a difficult task, particularly in trypanosomatids, where proteins
648 involved in the initiation of DNA replication are difficult to determine. Few proteins have been
649 described as potential ORC subunits (reviewed in [61]), and none of them have been shown to
650 be a specific marker that indicates the origins. We found a minority of TbORC1/CDC6 binding
651 sites near origins detected by SNS-seq and this was comparable with the previously reported
652 ~4.4% overlap of this protein with MFA-seq data (reviewed in [61]). Comparable results were also
653 obtained in *T. cruzi* [27]. DNAscent nanopore sequencing revealed 4,287 origins in *T. cruzi*. Of
654 these, 3,782 (88.2%) did not match the 851 TcORC1/CDC6-Ty1 binding sites within a ±3 kb
655 window [27]. Therefore, only 11.7% of origins of *T. cruzi* overlapped with TcORC1/CDC6-Ty1
656 binding sites within a ±3 kb window, a proportion similar to what we observed in *T. brucei*
657 (6.2%). It is also worth noting that origins detected by conventional SNS-seq in *Plasmodium*
658 *falciparum* revealed 63% of overlap with origins detected by nanopore sequencing and 13.6%
659 with $PfORC_{1-2}$ proteins [36]. Another group found no significant correlation between origins
660 detected by DNAscent nanopore sequencing and PfORC1 ChIP sites within the same organism
661 [62].

662 It is important to note that the topology of origins of *T. brucei* is very similar to that previously
663 described for other eukaryotes. The origins in mouse cells have been characterised as fragile
664 sites, defined by long poly (dA:dT) tracks that are nucleosome free and devoid of the single-
665 strand DNA-protecting protein RPA [13]. The same position and orientation of predicted G4
666 motifs, peaking at 240-300 base pairs from replication start sites, was previously identified in
667 two Drosophila cell lines using conventional SNS-seq [21], as was found here for *T. brucei*.
668 Furthermore, the organisation of nucleosomes around the origins of *T. brucei* is identical to
669 that observed in yeast and higher eukaryotes [12, 28-32, 63, 64]. In the case of *L. major* the origins
670 identified using conventional SNS-seq were also located in genomic regions with lower
671 nucleosome occupancy [22]. In addition, a significant correlation between *L. major* origins and
672 predicted G4 motifs was observed, with 74% of origins containing at least one such motif [22].
673 These similarities suggests that the structure of origins may be a conserved feature in all
674 eukaryotes. However, further studies are needed to gain a deeper insight into the evolution of
675 DNA replication origins.

676 **METHODS**

677 **Cell culture**

678 Procyclic forms (PCF) of *T. brucei*, strain Lister 427 wild type strain were grown at 27 °C in SDM-
679 79 (PAA Laboratories) supplemented with heat inactivated 10% FBS (Gibco) and 7 $\mu g.mL^{-1}$
680 hemin [65]. Bloodstream forms (BSF) of *T. brucei* Lister 427 strain were cultivated at 37°C, 5% CO

681 in a modified Iscove's culture medium (Gibco) complemented according to Hirumi *et al.* [66],
682 with 10% (v/v) heat-inactivated foetal bovine serum (Gibco), 36 mM sodium bicarbonate
683 (Sigma), 136 µg.mL^{-1} hypoxanthine (Sigma), 39 µg.mL^{-1} thymidine (Sigma), 110 µg.mL^{-1} Na-
684 pyruvate (Sigma), 28 µg.mL^{-1} bathocuproine (Sigma), 0.25 mM β-mercaptoethanol, 2 mM L-
685 cysteine.

686 **SNS isolation and purification**

687 SNS were isolated and purified following a previously published protocol [67] with few
688 modifications. Between 2 - 2.5x10^{9} cells were used as starting material for SNS isolation. Total
689 DNA was isolated from three biological replicates of asynchronous PCF and BSF cells in the
690 exponential phase of growth (3-5x10^{6} and 7-9x10^{5} cells.mL^{-1} for PCF and BSF respectively).
691 Cells were precipitated and incubated in 0.5 mL of 20 mM Tris-Cl, pH 7.0, 50 mM EDTA, 1% N-
692 Lauroylsarcosine, 0.5% Triton X-100 and 2 mg.mL^{-1} proteinase K and incubated at 45°C for 24
693 hours. Fresh solution of proteinase K was added after 24 hours and digested for 16 hours more.
694 DNA was purified with phenol/chloroform/isoamyl alcohol and precipitated in 0.3 M Na-
695 acetate (pH 5.3) and absolute ethanol. DNA was dissolved gently in TEN20 buffer (10 mM Tris-
696 Cl, 50 mM EDTA, 20 mM NaCl, pH 7.0) with 100 U RNasin (Promega). DNA was heat-denatured
697 (95°C/5 min) and loaded onto 5 mL of neutral 5%–20% sucrose gradients prepared in TEN20
698 buffer in ultra-clear tubes (Beckman Coulter). Size separation of nucleic acids was done by
699 ultra-centrifugation in SW50 rotor for 7 hours at 45 krpm at +4°C. Fractions of 200 µL from
700 the sucrose gradient were collected and checked by agarose gel electrophoresis to assess the
701 quality of nucleic acid separation (Supplementary Figure 10A). Fractions from 0.5-2.5 x10^{3} nt
702 were pooled together as SNS enriched fractions. Half of the SNS pool was treated with 10 µg
703 of cocktail of RNase, DNase free (Roche, Cat Nb 11119915001) for 16h at 37°C, to become the
704 SNS-depleted control. SNS-enriched samples, together with its SNS-depleted controls, were
705 treated 3-4 times with 50 U of T4 PNK (Thermo Fisher Scientific) and 100 U of λ-exonuclease
706 (Thermo Fisher Scientific). Each treatment was done after heat denaturation of DNA (95°C/10
707 min) and simultaneously on samples and controls. After denaturation, samples and controls
708 were immediately placed on ice. The enzymatic reactions were assembled on ice, and then the
709 enzymatic treatment with T4 PNK and λ-exonuclease was done at 37°C. After enzymatic
710 treatment and prior to library preparation, residual RNA was removed with 5 µg RNase, DNase-
711 free (Roche Cat Nb 11119915001), isolated DNA was sonicated with Bioruptor Pico sonication
712 device (Diagenode) and purified with magnetic beads (CleanNGS, Proteigen). The quantity of
713 purified material was estimated by fluorometer DS 11 FX (DeNovix) using Qubit ssDNA Assay
714 kit (Invitrogen).

715 It is important to note that we used SNS enrichment conditions and enzymatic treatments as
716 described in previous reports [20, 67]. We enriched the SNS by size on a sucrose gradient and
717 then treated this SNS-enriched fraction with high amounts of λ-exonuclease. A previous study
718 has shown that under these conditions, complete digestion of DNA containing G4-rich
719 sequences is achieved [20]. The efficiency of the T4 PNK and λ-exonuclease enzymes was tested
720 on both sheared gDNA and on RNA. The majority of DNA was digested after one treatment

721 with these two enzymes (Supplementary Figure 10B). RNA molecules treated under the same
722 conditions were preserved (Supplementary Figure 10C), suggesting that SNS would also be
723 preserved. The quantity of DNA molecules in SNS containing fractions was systematically
724 followed by qPCR after each enzymatic treatment and it was estimated that the amount of
725 DNA in starting material was reduced around 100 times after the first treatment, and around
726 10 times more after each subsequent treatment with these two enzymes.

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

728 Single-stranded SNS libraries were prepared using the Accel-NGS 1S Plus DNA Library Kit and
729 1S Plus Dual Indexing Kit for Illumina (Swift Biosciences), following the manufacturer's
730 protocol. A 3'end specific tail is preserving the strand orientation of the ssDNA fragments
731 during library preparation, resulting in a second strand library, where the plus-strand
732 fragments are represented by F1R2 read pairs and the minus-strand fragment by F2R1 read
733 pairs. The quality of the libraries was checked by capillary electrophoresis on Bioanalyzer 2100
734 (Agilent) using Agilent High Sensitivity DNA Kit (Agilent Technologies). The quantity of libraries
735 was estimated by fluorometry using dsDNA High Sensitivity Assay Kit on a DS-11 FX+
736 (DeNovix).

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

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

746 **SNS-seq data analysis**

747 Quality control of the FASTQ files was performed using FastQC (RRID:SCR_014583, v0.11.9)
748 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The sequencing reads were tail
749 trimmed to remove the 3'end specific tail introduced during library preparation using cutadapt
750 [68] with the following parameters (-u -10 -u 10 -U -10 -U 10). Afterwards the SNS-seq
751 sequencing reads were aligned against the *T. brucei* Lister 427-2018 reference genome release
752 55 from TriTrypDB (RRID:SCR_007043) (https://tritrypdb.org/common/downloads/release-
753 55/TbruceiLister427_2018, last accessed June 24th, 2025) with bowtie2 (RRID:SCR_016368 ,
754 v2.4.10) [69]. The alignment output bam files were converted, sorted, and indexed using samtools
755 [70]. Duplicated reads were removed using picard (RRID:SCR_006525, v2.23.5, "Picard Toolkit."
756 2019. Broad Institute, GitHub Repository. https://broadinstitute.github.io/picard/, last accessed
757 June 24th, 2025; Broad Institute). Mapped deduplicated reads pairs were then separated with
758 samtools (RRID:SCR_002105, v1.13) *view* into minus-strand and plus-strand read pairs via their

759 read orientation (minus-strand: F2R1 and plus-strand: F1R2). Peak calling was then performed
760 with macs2 (v.2.2.7.1) (using default parameters plus -p 5e-2 -s 130 -m 10 30 --gsize 2.5e7) for
761 plus and minus-strand reads separately with the SDC as control.

762 **Origin selection**

763 Origins were selected after the peak calling step, during the peak filtering step. The called
764 peaks were filtered using the bedtools suite (RRID:SCR_006646, v.2.30.0) [71] as described below.
765 In this peak filtering step, three criteria were defined to select the origins. The first criterion
766 was that two SNS peaks must be oriented in a way that reflects the topology of the SNS strands
767 within an active origin: the upstream peak must be on the minus strand and followed by the
768 downstream peak on the plus strand. The second criterion was to set a maximal inner to inner
769 border distance between two divergent minus and plus peak pair. The distances between the
770 start points of the two divergent leading strands are not known. We conducted tests on several
771 inner border-to-border distances, ranging from 0.5 to 2.5 kb. Higher numbers of origins were
772 found with larger distances. However, we finally applied a maximum distance of 0.5 kb,
773 presuming that this corresponds more closely to the physiological distance between two
774 divergent SNS molecules. The third criterion was introduced to prevent complete overlap of
775 two divergent peaks, as this can only occur if the two SNSs elongate at the same rate in both
776 directions, which is an unlikely scenario. Indeed, the synthesis of the Okazaki fragment and
777 ligation to the 5' of the SNS requires several enzymatic processes, and we can assume that the
778 elongation of SNS fibre is slower at the 5' end. Of note, complete overlap of minus and plus
779 peaks was rarely observed (mean 6.5% (stdev 6.5%, min 0.8%, max 15.8%)). To define the
780 maximum overlap between the minus and plus SNS peaks, a series of tests were conducted,
781 varying the percentage of overlap from 40% to 90%. This resulted in a similar number of
782 origins, and ultimately, 50% was selected as the optimal value. Therefore, firstly, minus-strand
783 and plus-strand peaks with an overlap of over 50% were removed with bedtools *intersect*.
784 Secondly peak pairs on different strands with a maximal distance of 0.5 kb were selected using
785 bedtools *windows*. Finally, peak pairs were filtered for the correct orientation (minus-strand
786 followed by plus-strand) with bedtools *closest* and the origin was defined as the genomic
787 region between the two inner borders of two divergent peaks. The overlap of origins between
788 replicates and cell types as well as other genomic features was analysed with bedtools *windows*
789 and bedtools *intersect*. Replicates were merged with bedtools *merge*.

790 For **visual inspection of genomic features** Integrative genomics viewer (IGV)
791 (RRID:SCR_011793) [72] was used. Violin plots were generated with the GraphPad Prism 8.3.0
792 software (RRID:SCR_002798, GraphPad Software Inc., La Jolla, CA, USA).

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

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

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

801 **PCF down-sampling** was performed to be able to compare origin number in BSF and PCF.
802 The mapped reads from all three PCF samples and control and BSF samples and control,
803 respectively, were merged using Samtools *merge*. The PCF combined sample was sub-samples
804 with Samtools *view* (--subsample 0.69 --subsample-seed [0-14]) and the merged control with
805 Samtools *view* (--subsample 0.57 --subsample-seed [0-14]) 15 times to obtain a set of mapped
806 reads comparable to the BSF combined sample. Origin detection was applied as described
807 above to the combined BSF and the 15 subsamples combined PCF samples.

808 For **visualization of the respective** localization of two or more datasets the the Deeptools
809 suite (RRID:SCR_016366, v3.5.1) [73] and the ggplot2 R package was used. The open source
810 ggplot2 R package (RRID:SCR_014601), freely available from the Comprehensive R Archive
811 Network (CRAN) repository (http://www.r-project.org, last accessed June 24th, 2025) was used
812 to generate composite plots. In Deeptools bed and bigwig files were used to generate a matrix
813 file with *ComputeMatrix* and *PlotHeatmap* or *PlotProfil* were used to generate visualizations
814 from the matrix.

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

819 **Nucleotide composition** around the origin centres was calculated as follows. The *T brucei*
820 genome was split into 20 bp windows using bedtools *window*. The number of A, T, G and C
821 nucleotides for each window were obtained with bedtools *nuc* and used to calculate the
822 nucleotide percentage per 20 bp window. The resulting bedgraph files were transformed into
823 bigwig file format using USCS utilities *bedGraphToBigWig*
824 (https://github.com/ucscGenomeBrowser/kent/tree/master/src/utils, last accessed June 24th,
825 2025).

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

827 **Handling of datasets mapped to different reference genomes.** The stranded SNS-seq data,
828 that we mapped to the Tb Lister 427-2018 genome, were compared with various published
829 datasets (SAS and PAS (TriTrypDB), MNase-seq [49], DRIP-seq [54], G4-seq [48], MFA-seq [42] and
830 Orc1/CDC6 ChIP-on-Chip [42]) that were analysed using different references genomes (Tb Lister
831 427 or TREU927). To enable direct comparison, we either reanalysed the published datasets
832 against the Tb Lister 427-2018 genome or remapped the stranded SNS-seq data to match the
833 respective reference genomes of the compared datasets. The mapping of origins in the

834 different reference genomes produced different numbers of origins, in Tb Lister 427 1059 PCF
835 and 1466 BSF origins and in total 2081 origins were detected, in TREU927 976 PCF and 1466
836 BSF origins were detected and in total 1963 origins were detected. Specific details on the
837 chosen approach are given in the relevant paragraphs below.

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

844 **MNase-seq.** The MNase-seq data [49] was downloaded as bigwig files from GEO (accession
845 number GSE90593). The data was published using the Tb Lister 427 reference genome. To be
846 able to compare our SNS-seq dataset with this data we reanalysed the SNS-seq sequencing
847 reads as described above in the Tb Lister 427 reference genome
848 (https://tritrypdb.org/common/downloads/release-62/TbruceiLister427/ last accessed June
849 24th, 2025). To quantify the MNase-seq signature we observed around the origin centres ("M"
850 pattern), we extracted the average MNase-seq signal (normalised dyad values [49]) from the
851 regions 400-200 nt upstream, 200-400 nt downstream and 75nt up- and downstream of the
852 origin centre with the UCSC utilities *bigWigAverageOverBed* command
853 (https://github.com/ucscGenomeBrowser/kent/tree/master/src/utils last accessed June 24th,
854 2025). Regions with negative values representing tandem repeats and N-rich sequences [49]
855 were discarded. LNO regions were considered as present when the average MNase-seq value
856 was below 0.23 normalised dyad value and the HNO regions were considered present when
857 the averaged normalised dyad value for the region was above this value. This value
858 corresponds to the average value for background regions (2000-1000 bp upstream of the
859 origin centres and 1000-2000 bp downstream of origin centres).

860 **Drip-seq.** The Drip-seq dataset was downloaded from ENA (accession number PRJEB21868)
861 [54] as fastq files. The re-analysis was done as described [54]. The published DRIP-seq analysis was
862 performed in Tb Lister 427 reference genome. We re-analysed the data in the Tb Lister 427
863 and the Tb Lister 427-2018 reference genomes. The intersection of DRIP-seq data with the
864 stranded SNS-seq origins was done in the Tb Lister 427-2018 reference genome.

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

870 **G4 prediction** in the *T. brucei* genome were done with G4Hunter [50] with the following settings,
871 window size (-w 25) and a threshold of 1.57. Prediction of G4 structures in the Tb 427-2018
872 reference genome and in the Tb Lister 427 reference genome were performed, to be able to

873 compare G4 structures to MNase-seq (in Tb Lister 427 reference genome) and DRIP-seq (in Tb
874 Lister 427-2018 reference genome) data as G4-seq output files were available only for the Tbb
875 TREU927 reference genome [48].

876 **MFA-seq.** We intersected six MFA-seq datasets. Two were obtained directly from the authors
877 [42] and four were downloaded from ENA (accession number PRJEB11437) [43], all in fastq file
878 format. The analysis of the fastq files was performed as described previously [42, 43]. The original
879 published MFA-seq analysis was performed in *T. brucei* Tbb TREU927 reference genome. We
880 re-analyzed the data in the Tbb TREU927 reference genome and in the Tb Lister 427-2018
881 reference genomes to be able to directly compare them to the stranded SNS-seq data.

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

890 **DNA molecular combing**

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

897 DNA molecular combing was done as described previously [41]. Protein-free plugs with DNA
898 were extensively washed in TEN20 buffer (10 mM Tris-Cl, 50 mM EDTA, 20 mM NaCl, pH 7.0)
899 and then stained with 0.5 µM YOYO-1 fluorescent dye (Molecular Probes) for 1 h at 37°C. After
900 washing from the excess YOYO-1 day, the 200 µL of TEN20 buffer was added to the plugs,
901 melted at 65 °C for 15 minutes, slowly cooled down to 42 °C and treated with 10 U of β-agarase
902 (Merck) for 16h. This treatment was repeated with fresh β-agarase for 4 hours. After digestion,
903 2 mL of 50 mM MES (mix of 0.5 M MES Hydrate and 0.5M MES Sodium Salt, pH 5.7) was added
904 very gently to the DNA solution and then DNA molecules were combed and regularly stretched
905 (2 kb/µm) on silanized coverslips provided by Montpellier DNA Combing Facility. Linearized
906 DNA molecules were fixed for 16 hours at 65 °C, denatured in 1N NaOH for 20 minutes and
907 blocked with 1 × PBS/ 1% BSA/ 0.1% Triton X-100. Immuno-detection was done with
908 antibodies diluted in 1 × PBS/ 1% BSA/ 0.1% Triton X-100 and incubated at 37 °C in a humid
909 chamber for 60 min. Each step of incubation with antibodies was followed by extensive washes
910 with 1 × PBS. Immuno-detection was done with anti-ssDNA antibody (1/100 dilution, clone 16-
911 19, Cat. No. MAB3034, Merck Milipore), the mouse anti-BrdU antibody (1/20 dilution, clone

912 B44, Cat. No. 347580, Becton Dickinson) and rat anti-BrdU antibody (1/20 dilution, clone
913 BU1/75, Cat. No. ABC117-7513, Abcys). The following secondary antibodies were used:
914 Alexa488 Goat anti Rat IgG (1/50 dilution, Cat. No. A-11006, Invitrogen), Alexa 546 Goat anti
915 Mouse IgG1 (1/50 dilution, Cat. No. A-21123, Invitrogen) and Alexa 647 goat anti-mouse IgG2a
916 (1/100 dilution, Cat. No. A-21241, Invitrogen). Coverslips were mounted with 15 µL of Prolong
917 Gold Antifade (Cat. No. P10144, Thermo Fisher Scientific) and dried at room temperature for
918 16 hrs. Images were acquired using Zeiss Z1 microscope, equipped with an ORCA-Flash 4.0
919 digital CMOS camera (Hamamatsu), using a 40× objective, where 1 pixel corresponds to 325 bp
920 (one microscope field of view corresponds to ~450 kb). Acquisition and analysis of images
921 were done by MetaMorph version 7 software (Molecular Devices LLC) (RRID:SCR_002368).

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

923 The speed of replication forks was estimated on individual forks, presented as IdU track
924 followed by a CldU track. Only intact (non-fragmented) forks were measured, which was
925 followed by DNA immuno-staining. The speed was calculated as the ratio between the CldU
926 track length (kilobases) and the CldU pulse duration (minutes). The IODs were measured as
927 the distance (in kilobases) between the centres of two adjacent progressing forks located on
928 an uninterrupted DNA fibre. Fork asymmetry was calculated as the ratio of the longer track
929 over the shorter track in progressing divergent forks. A ratio greater than one indicates
930 asymmetry. The GraphPad Prism 8.3.0 software (GraphPad Software Inc., La Jolla, CA, USA) was
931 used to generate graphs and to perform statistical analysis of DNA combing measurements as
932 described previously [41]. The statistical comparisons of different datasets were assessed by the
933 non-parametric Mann–Whitney two-tailed tests.

934 **DATA AVAILABILITY**

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

939 **AKNOWLEDGEMENT**

940 Special thanks to Cedric Notredame for advising us to prepare SNS stranded libraries. We would like to
941 thank Julie A.J. Clement and Stephanie Le Gras for their help with preliminary analysis of SNS-seq data.
942 We acknowledge the Montpellier DNA Combing Facility for providing silanized coverslips. We are
943 grateful to MGX-Montpellier GenomiX (Montpellier, France) facility for performing sequencing. We
944 would also like to acknowledge the 'Eukaryotic Pathogen, Vector and Host Informatics Resource'
945 (VEuPathDB) for the integrated database, covering numerous eukaryotic pathogens [76]. We are grateful
946 to the Roscoff Bioinformatics platform ABiMS (http://abims.sb-roscoff.fr), part of the Institut Français de
947 Bioinformatique (ANR-11-INBS-0013) and BioGenouest network, for providing computing resources.
948 Thanks to Frédéric Bringaud (University Victor Segalen Bordeaux 2) for providing the *T. brucei* 427 wild-
949 type PCF strain and Lucy Glover (Institute Pasteur, Paris) for providing us *T. brucei* 427 wildtype BSF
950 strain.

951 The work was supported by the Agence Nationale de la Recherche (ANR) within the frame of the
952 "Investissements d'avenir" program [ANR-11-LABX-0024-01 "PARAFRAP" to YS], the Centre National de
953 la Recherche Scientifique (CNRS), the French Ministry of Research and the Centre Hospitalier

Universitaire of Montpellier. SS was recipient of a grant from the Fondation pour la Recherche Médicale (ING20160435527). MGX acknowledges financial support from France Génomique National infrastructure, funded as part of "Investissement d'Avenir" program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09). ABiMS is founded as a part of the Institut Français de Bioinformatique (ANR-11-INBS-0013).

REFERENCES

1. Lee, C.S.K., Weiβ, M. & Hamperl, S. Where and when to start: Regulating DNA replication origin activity in eukaryotic genomes. *Nucleus* **14**, 2229642 (2023).
2. Hulke, M.L., Massey, D.J. & Koren, A. Genomic methods for measuring DNA replication dynamics. *Chromosome Res* **28**, 49-67 (2020).
3. Leonard, A.C. & Méchali, M. DNA replication origins. *Cold Spring Harb Perspect Biol* **5**, a010116 (2013).
4. Newlon, C.S. & Theis, J.F. The structure and function of yeast ARS elements. *Curr Opin Genet Dev* **3**, 752-758 (1993).
5. Okuno, Y., Satoh, H., Sekiguchi, M. & Masukata, H. Clustered adenine/thymine stretches are essential for function of a fission yeast replication origin. *Mol Cell Biol* **19**, 6699-6709 (1999).
6. Kong, D. & DePamphilis, M.L. Site-specific DNA binding of the Schizosaccharomyces pombe origin recognition complex is determined by the Orc4 subunit. *Mol Cell Biol* **21**, 8095-8103 (2001).
7. MacAlpine, D.M., Rodríguez, H.K. & Bell, S.P. Coordination of replication and transcription along a Drosophila chromosome. *Genes Dev* **18**, 3094-3105 (2004).
8. Balasov, M., Huijbregts, R.P. & Chesnokov, I. Role of the Orc6 protein in origin recognition complex-dependent DNA binding and replication in Drosophila melanogaster. *Mol Cell Biol* **27**, 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.e1119 (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.e4220 (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).

1003 20. Cayrou, C. *et al.* The chromatin environment shapes DNA replication origin organization and
1004 defines origin classes. *Genome Res* **25**, 1873-1885 (2015).
1005 21. Comoglio, F. *et al.* High-resolution profiling of Drosophila replication start sites reveals a DNA
1006 shape and chromatin signature of metazoan origins. *Cell Rep* **11**, 821-834 (2015).
1007 22. Lombraña, R. *et al.* Transcriptionally Driven DNA Replication Program of the Human Parasite
1008 Leishmania major. *Cell Rep* **16**, 1774-1786 (2016).
1009 23. Langley, A.R., Gräf, S., Smith, J.C. & Krude, T. Genome-wide identification and characterisation
1010 of human DNA replication origins by initiation site sequencing (ini-seq). *Nucleic Acids Res* **44**,
1011 10230-10247 (2016).
1012 24. Sequeira-Mendes, J. *et al.* Differences in firing efficiency, chromatin, and transcription underlie
1013 the developmental plasticity of the. *Genome Res* **29**, 784-797 (2019).
1014 25. Akerman, I. *et al.* A predictable conserved DNA base composition signature defines human
1015 core DNA replication origins. *Nat Commun* **11**, 4826 (2020).
1016 26. Prorok, P. *et al.* Involvement of G-quadruplex regions in mammalian replication origin activity.
1017 *Nat Commun* **10**, 3274 (2019).
1018 27. Vitarelli, M.d.O. *et al.* Integrating high-throughput analysis to create an atlas of replication
1019 origins in Trypanosoma cruzi in the context of genome structure and variability. *mBio* **15**,
1020 e00319-00324 (2024).
1021 28. Lipford, J.R. & Bell, S.P. Nucleosomes positioned by ORC facilitate the initiation of DNA
1022 replication. *Mol Cell* **7**, 21-30 (2001).
1023 29. Berbenetz, N.M., Nislow, C. & Brown, G.W. Diversity of eukaryotic DNA replication origins
1024 revealed by genome-wide analysis of chromatin structure. *PLoS Genet* **6**, e1001092 (2010).
1025 30. Foss, E.J. *et al.* Identification of 1600 replication origins in S. c*erevisiae*. *Elife* **12**, RP88087
1026 (2024).
1027 31. Poulet-Benedetti, J. *et al.* Dimeric G-quadruplex motifs-induced NFRs determine strong
1028 replication origins in vertebrates. *Nat Commun* **14**, 4843 (2023).
1029 32. Jansen, A. & Verstrepen, K.J. Nucleosome positioning in Saccharomyces cerevisiae. *Microbiol*
1030 *Mol Biol Rev* **75**, 301-320 (2011).
1031 33. Nasheuer, H.P. & Meaney, A.M. Starting DNA Synthesis: Initiation Processes during the
1032 Replication of Chromosomal DNA in Humans. *Genes* **15**, 360 (2024).
1033 34. Bielinsky, A.K. & Gerbi, S.A. Discrete start sites for DNA synthesis in the yeast ARS1 origin.
1034 *Science* **279**, 95-98 (1998).
1035 35. Foulk, M.S., Urban, J.M., Casella, C. & Gerbi, S.A. Characterizing and controlling intrinsic biases
1036 of lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes
1037 and G-quadruplex motifs around a subset of human replication origins. *Genome Res* **25**, 725-
1038 735 (2015).
1039 36. Castellano, C.M. *et al.* The genetic landscape of origins of replication in P. falciparum. *Nucleic*
1040 *Acids Res* **52**, 660-676 (2024).
1041 37. Senkevich, T.G. *et al.* Mapping vaccinia virus DNA replication origins at nucleotide level by deep
1042 sequencing. *Proc Natl Acad Sci U S A* **112**, 10908-10913 (2015).
1043 38. Fenn, K. & Matthews, K.R. The cell biology of Trypanosoma brucei differentiation. *Curr Opin*
1044 *Microbiol* **10**, 539-546 (2007).
1045 39. Baldauf, S.L. The deep roots of eukaryotes. *Science* **300**, 1703-1706 (2003).
1046 40. Clayton, C. Regulation of gene expression in trypanosomatids: living with polycistronic
1047 transcription. *Open Biol* **9**, 190072 (2019).
1048 41. Stanojcic, S. *et al.* Single-molecule analysis of DNA replication reveals novel features in the
1049 divergent eukaryotes *Leishmania* and *Trypanosoma brucei* versus mammalian cells. *Sci Rep* **6**,
1050 23142 (2016).
1051 42. Tiengwe, C. *et al.* Genome-wide analysis reveals extensive functional interaction between DNA
1052 replication initiation and transcription in the genome of Trypanosoma brucei. *Cell Rep* **2**, 185-
1053 197 (2012).

1054 43. Devlin, R. *et al.* Mapping replication dynamics in Trypanosoma brucei reveals a link with
1055 telomere transcription and antigenic variation. *Elife* **5**, e12765 (2016).
1056 44. Fragkos, M., Ganier, O., Coulombe, P. & Mechali, M. DNA replication origin activation in space
1057 and time. *Nat Rev Mol Cell Biol* **16**, 360-374 (2015).
1058 45. Techer, H. *et al.* Replication dynamics: biases and robustness of DNA fiber analysis. *J Mol Biol*
1059 **425**, 4845-4855 (2013).
1060 46. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo generator.
1061 *Genome Res* **14**, 1188-1190 (2004).
1062 47. Valton, A.L. & Prioleau, M.N. G-Quadruplexes in DNA Replication: A Problem or a Necessity?
1063 *Trends Genet* **32**, 697-706 (2016).
1064 48. Marsico, G. *et al.* Whole genome experimental maps of DNA G-quadruplexes in multiple
1065 species. *Nucleic Acids Res* **47**, 3862-3874 (2019).
1066 49. Maree, J.P., Povelones, M.L., Clark, D.J., Rudenko, G. & Patterton, H.G. Well-positioned
1067 nucleosomes punctuate polycistronic pol II transcription units and flank silent. *Epigenetics*
1068 *Chromatin* **10**, 14 (2017).
1069 50. Brázda, V. *et al.* G4Hunter web application: a web server for G-quadruplex prediction.
1070 *Bioinformatics* **35**, 3493-3495 (2019).
1071 51. Sollier, J. & Cimprich, K.A. Breaking bad: R-loops and genome integrity. *Trends Cell Biol* **25**, 514-
1072 522 (2015).
1073 52. Boguslawski, S.J. *et al.* Characterization of monoclonal antibody to DNA.RNA and its
1074 application to immunodetection of hybrids. *J Immunol Methods* **89**, 123-130 (1986).
1075 53. Wahba, L., Costantino, L., Tan, F.J., Zimmer, A. & Koshland, D. S1-DRIP-seq identifies high
1076 expression and polyA tracts as major contributors to R-loop formation. *Genes Dev* **30**, 1327-
1077 1338 (2016).
1078 54. Briggs, E., Hamilton, G., Crouch, K., Lapsley, C. & McCulloch, R. Genome-wide mapping reveals
1079 conserved and diverged R-loop activities in the unusual genetic landscape of the African
1080 trypanosome genome. *Nucleic Acids Res* **46**, 11789-11805 (2018).
1081 55. Godoy, P.D. *et al.* Trypanosome prereplication machinery contains a single functional
1082 orc1/cdc6 protein, which is typical of archaea. *Eukaryot Cell* **8**, 1592-1603 (2009).
1083 56. Zerbib, A. & Simon, I. Characterization of Unidirectional Replication Forks in the Mouse
1084 Genome. *Int J Mol Sci* **24**, 9611 (2023).
1085 57. Bou-Nader, C., Bothra, A., Garboczi, D.N., Leppla, S.H. & Zhang, J. Structural basis of R-loop
1086 recognition by the S9.6 monoclonal antibody. *Nat Commun* **13**, 1641 (2022).
1087 58. Hyrien, O. Peaks cloaked in the mist: the landscape of mammalian replication origins. *J Cell*
1088 *Biol* **208**, 147-160 (2015).
1089 59. Tian, M. *et al.* Integrative analysis of DNA replication origins and ORC-/MCM-binding sites in
1090 human cells reveals a lack of overlap. *eLife* **12**, RP89548 (2024).
1091 60. Marques, C.A., Dickens, N.J., Paape, D., Campbell, S.J. & McCulloch, R. Genome-wide mapping
1092 reveals single-origin chromosome replication in *Leishmania*, a eukaryotic microbe. *Genome*
1093 *Biol* **16**, 230 (2015).
1094 61. Marques, C.A. & McCulloch, R. Conservation and Variation in Strategies for DNA Replication of
1095 Kinetoplastid Nuclear Genomes. *Curr Genomics* **19**, 98-109 (2018).
1096 62. Totañes, F.I.G. *et al.* A genome-wide map of DNA replication at single-molecule resolution in
1097 the malaria parasite Plasmodium falciparum. *Nucleic Acids Res* **51**, 2709-2724 (2023).
1098 63. Segal, E. & Widom, J. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr*
1099 *Opin Struct Biol* **19**, 65-71 (2009).
1100 64. Xu, J. *et al.* Genome-wide identification and characterization of replication origins by deep
1101 sequencing. *Genome Biol* **13**, R27 (2012).
1102 65. Wirtz, E., Leal, S., Ochatt, C. & Cross, G.A. A tightly regulated inducible expression system for
1103 conditional gene knock-outs and dominant-negative genetics in Trypanosoma brucei. *Mol*
1104 *Biochem Parasitol* **99**, 89-101 (1999).

66. 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).
67. Cayrou, C., Grégoire, D., Coulombe, P., Danis, E. & Méchali, M. Genome-scale identification of active DNA replication origins. *Methods* **57**, 158-164 (2012).
68. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10-12 (2011).
69. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **35**, 421-432 (2019).
70. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
71. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
72. Robinson, J.T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26 (2011).
73. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165 (2016).
74. 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).
75. 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).
76. Amos, B. *et al.* VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Res* **50**, D898-D911 (2022).