Supplementary Information:

Isolation and sequencing of active origins of DNA replication by nascent strand capture and release (NSCR)

Dimiter Kunnev[1], Amy Freeland[1], Maochun Qin[2], Jianmin Wang[2], and Steven C. Pruitt[1]

[1]*Department of Molecular and Cellular Biology*

[2]*Department of Biostatistics and Bioinformatics*

Roswell Park Cancer Institute, Buffalo, New York 14263, USA

Corresponding author: steven.pruitt@roswellpark.org

Table of contents:

I. Supplementary introduction figure

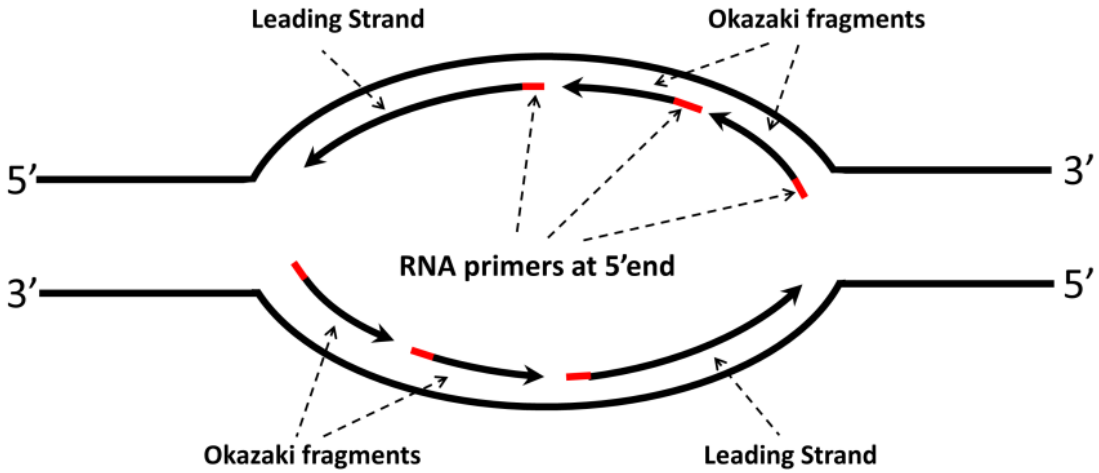

Figure S1. Schematic view of a replication origin.
DNA dependent DNA polymerases extend in a 5' to 3' direction and require a RNA primer to start the DNA polymerization process. The primer is synthesized by primase activity in slightly different length among the species (for yeast 8-10nt, for Drosophila 8-15, for mouse 9-11nt, and for Human 11-14nt) [1]. The resulting nascent strand then consists of a chimeric RNA-DNA molecule. On one side of each replication fork DNA is extending 5' to 3' in the direction of fork progression and this strand is termed the leading strand. Conversely, on the other side of replication fork, 5' to 3' extension of DNA on the "lagging strand" is opposite to the direction of fork movement and the synthesis proceeds via 200-300 nucleotide fragments called Okazaki fragments. Those fragments are ligated following removal of the RNA primer and gap repair. Okazaki fragments are present at all locations where DNA replication forks are extending, which may be at considerable distances from the origin. Up until the point that a nascent strand is ligated to a strand originating from an adjacent origin, it contains an RNA primer at the 5' end.

II. Supplementary methods

A. Purification of genomic DNA from cultured cells by SDS Lysis Buffer and Protease K (TIMING: 1 Day)

1. Wash ~$1.5x10^8$ - $2x10^8$ cultured cells with PBS.

2. Add lysis buffer with Proteinase K directly to the cells for example 4ml per plate, swirl well. Alternatively, if you are harvesting suspension cells add 50ml lysis buffer with Proteinase K per sample, re-suspend well and avoid forming clumps.

3. Split or combine the lysate from step 2 into 50ml Falcon tubes with max volume of 25ml per tube. Use as many Falcon tubes as needed.

4. Incubate at 37ºC for 16h (Alternately 2-3h at 55ºC).

5. Add an equal volume of phenol (pH 8.0), tightly close the tube, and swirl gently for one hour or more.

6. Centrifuge at 3000xg for 30 min, collect supernatant with pre-cut 1ml tip or with 5ml pipet.

7. Repeat steps 5 and 6 using chloroform-isoamyl alchohol.

8. Add an equal volume of isopropanol invert the tube several times until you observe the DNA aggregates, incubate at -20ºC for 10 minutes or longer and centrifuge at 3000 x g for 30 minutes to pellet the DNA.

9. Wash well with at least 10ml 70% ethanol.

PAUSE POINT: If you need you can stop since it is better to preserve the DNA with 70% ethanol

10. Air dry the pellet, add 5ml TE buffer and dissolve it well. If it is necessary you may incubate on ice until next day.

TIP: To avoid unnecessary damage to DNA, pipetting the DNA lysate needs to be minimized.

NOTE: DNA from tissue may require additional initial steps including homogenization and protein precipitation which current protocol does not describe.

B. Separation of DNA from RNA by using TRIzol Reagent: Example for 5ml dissolved DNA (TIMING 1 Day)

1. In a glass centrifuge tube mix 5ml DNA dissolved in TE from step 10 with 12.85ml TRIzol and 3.75ml chloroform. Cap the tube tightly; shake vigorously and incubate for 2-15 minutes at room temperature.

CAUTION: In order to avoid a spill from the glass tube, you can shake the TRIzol mixture in a capped Falcon tube first and then to transfer it into the glass tube. The recommendation from TRIzol guide-sheet is to centrifuge the mixture at 12000 x g. The glass tubes can tolerate higher G force more than 3000 x g therefore it need to be performed into glass tubes if you would like to follow the recommended G force. However, we obtained also a good separation using Falcon tubes at 2900 x g but centrifuging for a longer time 45 minutes.

2. Centrifuge the sample in glass tubes at 12000 x g for 15 minutes at 4ºC or alternatively if you use Falcon tubes at 2900 x g for 45 minutes.

3. Remove the top layer (containing the RNA), gently pick up the white interphase (containing the DNA) by inserting a cut pipet tip into the interphase above the phenol (red) phase and place it in a 15ml falcon tube.
NOTE: the top layer contains mostly RNA and is not viscous; however the interphase is white and rubber-like in composition. This layer is mostly genomic DNA and requires special treatment to dissolve.

CAUTION: Do not follow the rest of the TRIzol protocol provided within kit as it uses from the company for treatment with NaOH in order to dissolve the DNA. Alkali conditions will destroy the RNA primers required for SNS capture.

4. Wash the DNA several times with 14ml TE (usually 3-5 times). The DNA will NOT dissolve yet.

5. Add 5ml TE and incubate on ice overnight until DNA is dissolved well. Add 5ml isopropanol to re-precipitate DNA. Centrifuge it again 3000 x g for 30 min. Wash the pellet with 70% ethanol. You can pause for longer time with 70% ethanol.

NOTE: if DNA has not dissolved after the overnight incubation heat it at 55ºC for ~2-3h or until it is well dissolved. Traces of TRIzol will remain at this point and will interfere with absorbance at 260/280nm.
PAUSE POINT: DNA can be stored under 70% ethanol at -20 ºC for long period of time.

6. Dry the pellet and dissolve it into 1ml TE buffer. The second time the DNA will dissolve easily. Run a small aliquot into a 1% agarose gel with and without RNase A to assess the level of RNA contamination. Some contaminating RNA may be present. Due to the fact that this DNA was not heated (not denatured) yet, it will stay as very high band above 20 kb, the RNA will be seen if any as smear up to 2 kb. If minimal compared to the DNA it is not a concern. However, if the level of RNA equals or exceeds the amount of DNA it may reduce the yield of nascent strand in later steps.

7. Heat the DNA in TE at 95-100ºC and immediately transfer it to ice. Measure the concentration of DNA by absorbance at 260nm. A typical expected yield is ~1-1.5mg total DNA per sample, but depending on the initial source the amount may vary. If you are comparing different samples, normalize the concentration using TE buffer.
PAUSE POINT: DNA can be stored at -20 ºC or could be loaded on the sucrose gradient for size fractionation.

TIP: DNA must be sufficiently denatured to separate the genomic DNA from nascent DNA strands resulting from replication. $A_{260/280}$ measurements are more easily made at this step since the ssDNA is reduced in viscosity.

C. Preparation of sucrose gradient manually: (TIMING 15 minutes per gradient)

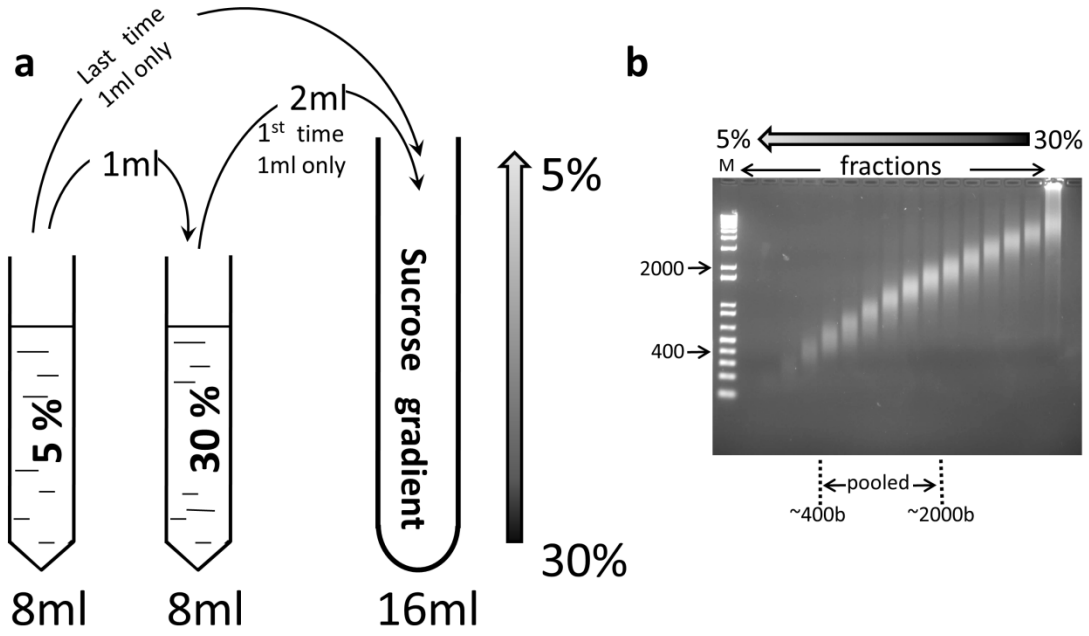

Figure S2. Preparation of 5% - 30% sucrose gradients manually (optional).
The example is for preparation of a 16ml gradient in 17ml SW28.1 Backman rotor tubes. a) Schematic representation of gradient preparation. b) Example of agarose gel electrophoresis of sucrose gradient fractionated DNA. Volumes can be scaled if different size tubes/alternative rotors are used.

STEPS:

1. Prepare two unused falcon tubes; one with 30% sucrose buffer and the other with 5% sucrose buffer, 8ml each. In addition, prepare a clean empty 17 ml gradient tube for the SW28.1 rotor next to the falcon tubes.
2. Transfer 1ml buffer with a cut tip from the 30% falcon tube to the bottom of the gradient tube.
3. Transfer 1ml buffer from 5% falcon tube to the 30% falcon tube.
4. Vortex the 30% falcon tube.
5. Transfer 2ml buffer from well vortexed 30% falcon tube to the gradient tube by gently lay it on the previously added buffer.

TIP: Each time 1ml of 5% buffer is transferred to the 30% tube, vortex thoroughly. It is also important to gently overlay the 2ml aliquots of sucrose buffer onto the surface of the previously loaded buffer into gradient tube and avoid extensive mixing.

6. Repeat steps 3 to 5 until only ~1ml is left in the 5% falcon tube and nothing remains in the 30% tube.
7. Transfer the last 1ml buffer from 5% tube to the top of the gradient to complete the process.

TIP: You can keep the gradients in a rack covered with parafilm at 4ºC for an hour or two until you finish with all of the gradients.

D. Size fractionation of nascent DNA + DNA fragments by 16ml 5%-30% sucrose gradient (TIMING 1 Day)

1. Prepare the required number of 16ml (5-30%) sucrose gradients with gradient maker or manually as it is described in **Figure S2**. Equalize the weights of all gradients by slightly adding or removing 5% sucrose buffer to or from the top.

2. Overlay (load) the purified and heat denatured DNA in TE (purification of DNA prior is described in supplementary information) on the top of the gradients (max ~400µg per gradient). Use at least three gradients per sample.

3. Ultracentrifuge the gradients at 4ºC/26000 rpm for 20h.

4. Fractionate the gradient into 16 fractions. Collect 1ml fractions by pipetting 1ml aliquots from the top of the gradient and place the aliquots in separate tubes.
PAUSE POINT: The fractions can be stored at -20 ºC up to 10 days, if need to store for a longer time keep it at -80 ºC.

5. Run 30µl from each aliquot on a 1% agarose gel to estimate the size of DNA in each fraction.

6. Collect and pool the desired size fractions.

NOTE: Typically, DNA in the top 3-4 fractions (smallest) cannot be seen on the gel, but a graded increase in size from the top to the bottom is observed in the remaining fractions (**Fig. S2b**). We collect and pool fractions from 4 to 9 corresponding to ~400-2000 base DNA fragments; however, if you run different volume gradients the migration of DNA fragments may differ. Once conditions are well established, the agarose gel can be omitted.

7. Add glycogen to 1µg/ml and precipitate the pooled DNA fractions with 2.5 volumes of ethanol or 1 volume of isopropanol with incubation at -20ºC.
PAUSE POINT: DNA can be stored for long periods of time under ethanol or isopropanol at -20ºC.

CAUTION: If you collect fractions from the sucrose gradient containing higher than ~10-15% of sucrose, you need to dilute the pooled fractions with RNase/DNase free ddH_2O in order to reduce the sucrose concentration prior to adding alcohol.

8. Centrifuge the samples at 3000 x g for 45 minutes (for falcon plastic tubes). If you use glass tubes, the yield could be improved by spinning faster (e.g. 10,000 x g).

9. Wash well with 70% ethanol.
PAUSE POINT: DNA can be stored for a long time under 70% ethanol at -20ºC.

10. Centrifuge and dry the pellets, re-suspend them in 50-60µl or larger volume of TE, and measure absorbance at 260/280nm. A typical yield is ~100-200 µg of sized pooled DNA from starting 1mg RNA free DNA. Proceed directly with biotinylation or keep it at -20°C for up to one week.

E. 5'-labeling nascent strand DNA + DNA fragments with biotin for one reaction (for up to 0.6 nmol 5' ends - can be scaled for larger amounts) (TIMING 4h)

CAUTION: use a formula for proper calculation of your nmols 5'ends

$$\frac{A}{333 \times C} \times 1000 \text{ nmols}/\mu\text{mol} = \text{nmols of 5' ends per } \mu\text{l}$$

- A = Conc of your ssDNA per µl (µg/µl)
- 333 is average mol weight of one nucleotide (ssDNA)
- C= average length of your fragments: for nascent DNA this will be between a minimum of 400 and a maximum of 1500 to 2000 nt.

Your result will be in nmols per µl. Multiply by the volume to determine the total nmols of 5' ends. Example for lower size 400nt and for 100 µg with concentration ~1 µg/µl it will be ~0.75 nmols. For higher size 2000nt and same amount and concentration will be 0.15 nmols. Therefore 100 µg pooled 400-2000nt fractions would be averaged as ~ 0.6 nmols, exactly as we need for one reaction. However, a concentration as 100µg of DNA in 2 µl is unrealistic to achieve, therefore we are performing 5'-biotinylation with similar amount in a 10x or 20x volumes scaled up reaction. The following steps are example for starting volume of 8µl.

1. Make a 10µl reaction in an Eppendorf tube with 1µl universal buffer, 1µl phosphatase (CIP), a maximum of 0.6 nmol of 5'-ends from step 10 (described in Size fractionation of nascent DNA + DNA fragments) and add ddH_2O to bring to volume.
2. Incubate at 37ºC for 1h.

3. Add 2µl universal buffer, 2µl ATPγS, 2µl T4 polynucleotide kinase and 4 µl ddH_2O to bring the volume to 20 µl .

4. Incubate at 37ºC for 1h.

5. Add 10µl of biotin maleimide dissolved in DMF.

6. Incubate at 65ºC for 1-2h.

7. Add 500µl TE to the reaction at room temperature.

8. Add 100µl phenol and shake well.

9. Centrifuge 14,000 rpm (max rpm) for 10 min.

10. Collect the aqueous (upper) phase (~500µl) into a new tube.

11. Add 200µl TE to the phenol to re-extract.

12. Repeat steps 9 and 10.

13. Combine first and second aqueous phases (~700µl total).

CAUTION: The reduced volume of phenol (less than 1:1) for extraction and additional re-extraction is designed to improve the yield of the biotinylated nucleic acid.

14. Add 5µl glycogen and 50µl 3M NaAc (or use the precipitant provided in the Vector Laboratories kit: Cat. MB-9001 kit).

15. Add 700µl isopropanol and incubate at least 30 minutes
PAUSE POINT: If you need you can keep the samples longer at -20ºC

16. Spin in microfuge at maximum rpm for 45 min.

17. Remove supernatant taking care not to disturb the pellet.

18. Wash with 70% ethanol three times. Be careful not to lose the pellet, centrifuge 5 minutes at maximum rpm for 5 minutes between every wash.
PAUSE POINT: You may store the samples in 70% ethanol at -20ºC for a longer period of time.

19. Dry the pellet and dissolve it in 50µl TE.

20. Measure the absorbance at 260/280nm and calculate the amount of nucleic acid

F. Ribonuclease I (RNase I) digest single incorporated ribonucleotide into ssDNA

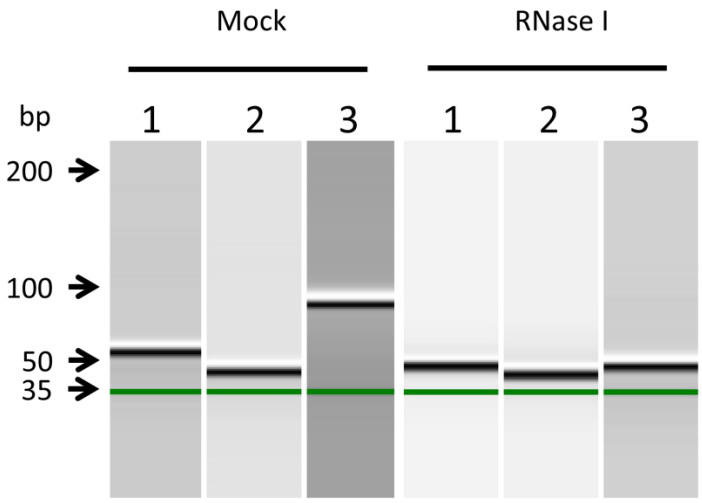

Figure S3. Activity of RNase I on a deoxy-oligonucleotide with a single incorporated ribonucleotide.

To test the ability for RNase I to cleave a single ribonucleotide flanked by ssDNA we used 3 oligos synthesized by Integrated DNA Technologies (IDT):

Oligo 1 (12rNTP-20dNTP):

rArUrCrArGrArUrCrArCrUrUCAGCTCCTCGGCTTAAAATG ;

Oligo 2 (12rNTP-20dNTP):

rGrGrUrGrCrArUrUrUrCrUrGTCCTCGCAGTAACGCAATAG

Oligo 3 (31dNTP-1rUTP-30dNTP):

TACCCACACTCCTGCTTGACCACCTTGTATG**rU**ACTGATTCCCCTACCATCCTC ACACCAACA

The oligonucleotides shown above were incubated in RNB buffer in the absence (mock) or presence (RNase I) of RNase I (0.02U/µl) for 15 minutes at room temperature. Purified samples were run on a High Sensitivity DNA Assay bioanalyzer chip to detect size differences. Oligos 1 and 2 are shorted by RNase I digestion as expected. Oligo 3 is reduced to approximately half of its starting size consistent with cleavage at the single rUTP in the middle of the ssDNA oligo. Discrepancies between the migration of the double strand size standards (at left of figure) and the predicted migration of the single stranded oligonucleotides may result from secondary structure.

III. Bioinformatics and analysis of nascent strand sequences

A. Software requirement

A list of public available software used for NSCR data analysis is listed below:

1. bwa 0.6.2 : sequence mapping
2. samtools 0.1.19: bam file operation
3. Picard tools : mark duplicate
4. MACS 1.4.2: wiggle file generation and peak calling
5. bedtools 2.17.0: bed file comparison
6. MEME/DREME 4.9.0 : sequence motif search

We also developed some custom scripts and programs for the data analysis. They can be downloaded at: http://sourceforge.net/projects/nscr/files/

NOTE: JAVA®, R and Perl should be installed prior to use those programs.

1. peak.calling.R : a simple custom peak calling R script
2. reformatWig.pl: convert wiggle file to peak.calling.R input file
3. WigSubtractor.jar : a java program to do wiggle file subtraction
4. RetrievePeakSequence.jar : a java program to extract sequences around peaks/origins
5. motifSearch.pl : searching for motifs
6. calculateBendability.pl : calculate DNA bendability
7. calculateStiffness.pl : calculate DNA stiffness

All softwares are installed under CentOS 5 with kernel version of 2.6.18 on a high performance computer cluster and a local workstation with Ubuntu 14.04.

B. Hardware requirement

All of the analyses are run on a high performance computing (HPC) cluster with each node having 4 Intel® Xeon® E5-2670 processors, 32G of memory, and 1TB of hard drive. For users without access to HPC, a workstation with enough power to handle bwa reads mapping will also work. For example, our workstation with 2 Intel ® Xeon ® E5-2620 @ 2.10GHz processors, 32G of memory, and 1TB of hard drive can perform the data analysis.

C. Mapping reads

High quality paired-end reads passing the Illumina RTA filter are aligned to the NCBI mouse reference genome using BWA version 0.6.2 [2]. The generated BAM files were further sorted and indexed using SAMtools version 1.0 [3]. PCR duplicated reads were marked and removed using Picard tools (http://picard.sourceforge.net/). We used mm9 reference genome file and the file name is `mm9.fa`. The commands for this step is listed below:

1.index the reference genome file with samtools

```
bwa index mm9.fa
```

This command will generate `mm9.fa.fai` index file.

2. index the reference genome file with bwa

```
bwa index mm9.fa
```

This command will generate `mm9.fa.amb, mm9.fa.ann, mm9.fa.bwt, mm9.fa.pac`, and `mm9.fa.sai` for bwa mapping.

3. create dictionary file for picard

```
java -jar CreateSequenceDictionary.jar R=mm9.fa
     O=mm9.dict
```

This command will generate `mm9.fa.dict` file.

4. bwa mapping for pair-end reads file `read_1.fastq` and `read_2.fastq`

```
bwa aln mm9.fa read_1.fastq > read_1.sai
bwa aln mm9.fa read_2.fastq > read_2.sai
bwa sampe mm9.fa read_1.sai read_2.sai read_1.fastq
     read_2.fastq > aln.sam
```

This will generate `aln.sam` file.

4. Convert sam to bam, sort, index, mark duplication and take flagstat for bam file.

```
samtools view -bS aln.sam > aln.bam
samtools sort aln.sorted.bam aln.bam
samtools index aln.sorted.bam
java -jar Markdup_compiled.jar INPUT=input.bam \
     OUTPUT=input.markdup.bam TMP_DIR=./  \
     METRICS_FILE=duplicate_metrics \
     REMOVE_DUPLICATES=false ASSUME_SORTED=true \
     VALIDATION_STRINGENCY=SILENT CREATE_INDEX=true
Samtools flagstat aln.sorted.markdup.bam
```

After those, a bam file named aln.sorted.bam and its index file aln.sorted.bam.bai will be created and the bam file has duplicate reads marked. For our 129 wild type data, the flagstat result is like:

```
395601890 + 0 in total (QC-passed reads + QC-failed
reads)
0 + 0 duplicates
343724073 + 0 mapped (86.89%:nan%)
395601890 + 0 paired in sequencing
197800945 + 0 read1
197800945 + 0 read2
292429292 + 0 properly paired (73.92%:nan%)
309716804 + 0 with itself and mate mapped
34007269 + 0 singletons (8.60%:nan%)
13063734 + 0 with mate mapped to a different chr
8283935 + 0 with mate mapped to a different chr
(mapQ>=5)
```

D. Wiggle file generation

Wiggle files are generated using MACS version 1.4 [4] using the following parameters: `tag size=50` and `band width=250`. The command is:

```
macs14 -t aln.bam -f BAM -g hs -n sample_name –nomodel \
      --petdist=250 --shiftsize=50 -p 0.001 –w
```

This command will generate wiggle files under `./treat` directory for each chromosome with wig extension. This command can also generate peak files. All wiggle files are further compressed into gz files to reduce file sizes, such that they can be easily loaded into the UCSC genome browser [5] for visualization purposes.

E. Peak calling/putative replication origin identification

Peaks can be detected using existing tools such as MACS or Homer [6], which are designed to identify transcription factor binding sites of ChIP-Seq experiments. The command to generate wiggle files using MACS will also identify enriched peaks. Due to the nature of nascent strand sequencing, such as potential wider peaks, multiple overlapping peaks in a small window (**Figure S4** shows this common feature in our data), most available peak detection methods are not designed for this purpose. Here we implemented a simple method to predict locations of peaks representing potential replication origins as described below.

E.1 Selection of windows with potential peaks

Windows, that were covered with minimum read depth and spanning a minimum width, are selected for peak identification. We set the default minimum widow width in our data analysis as 300 due, first, to the selected size of nascent strand DNAs (400 to 2000 nt) and, second, the average length library insert of 250 bp. We set default minimum read depth cutoff value as 1, which means any window with reads will be considered and no covered peak should be missed for peak detection.

E.2 Peak detection in selected windows

Peak detection depends on whether a control input dataset is available. Although not optimal, when no input data, the empirical coverage distribution of current chromosome from test dataset will be treated as the background and a cutoff value for coverage will be determined depending on the significance level. We selected the 95 percentile of the background distribution for the default cutoff value. User can specify the cutoff value by either percentile or by an absolute peak height. If the maximum height of a window exceeds this cutoff value, this window contains one or more peaks.

Use of input data to control for sequencing and other biases is preferred where one useful source of DNA for generation of these data is genomic DNA from which short nascent strands have been stripped and which is generated during the sucrose gradient fractionation step in the present protocol. Ideally, this input DNA will be sheared and treated in parallel with NSCR-SNS DNA including sequencing on the same sequencing lane using tagged libraries. When input data is available, the same window will be extracted from the input data. Many enrich tests can be used to test whether the window contains significant more reads or the coverage is significant higher in test data. We used Kolmogorov–Smirnov one sided test on the coverage distribution. Alternatively, for

reads count data, Fisher's exact test can be employed. The significant level can be specified by the user with default value of 1×10^{-5}. We set this small value to compensate multiple test problems.

E.3 Exact replication origin identification

As mentioned before, in NSCR data, wide windows with overlapping peaks are common, for any window that has peak detected in previous step, we identify the exact origin using the following method. A smoothing spline is fitted to the coverage data using smooth.spline R function with the degree of freedom parameter set to current window width divided by 100, this parameter can be specified by the user. The maximum values of the smoothed spline are summits of peaks and could be the replication origins. The global maximum is detected as a potential origin. For each local maximum, it is identified as a origin if the following three criteria are met: 1) the summit coverage must exceed the minimum peak height defined as previously; 2) the distance between two peaks should be greater than a cutoff value (we set 200bp as the default value as two peaks close to each other will merge into a single peak in coverage data); and 3) the predicted coverage drop, which is calculated by the difference of predicted summit height and valley height using smoothed spline model between two peaks should be greater than a cutoff value (e.g. in our experiments [11] a cutoff value is maximum 10 and 10% of peak heights and is based on the average peak heights 20). An example of two peaks identified in a window is shown in **Figure S4a**.

For the wild type 129S mouse without input data, our method generates 625,830 potential peaks using 95 percentile as cutoff value. Using MACS without building the shifting model (--nomodel parameter), 112,307 potential peaks are identified, and among those, only two peaks are not detected by our method, those two peaks have lower maximum peak height than the 95 percentile cutoff value. An example of peak identified in our approach but missed by MACS is shown in **Figure S4b** which is similar to **Figure S4a** but with only one peak. This specific peak resides in a 3kb region which is wider than most transcription binding footage also it has multiple potential peaks in it. The potential pitfalls of our approach include the high sensitivity may cause low specificity and reads from repeat region could be called as peaks without input dataset.

NOTE: Systematically estimation of the sensitivity and specificity of our replication origin identification requires a known true dataset with every origin identified and verified which is absent at this time. Such dataset would be of great value for evaluating our protocol in replication origin detection.

Running this code requires knowledge of R programming language and the following example shows how to do peak calling on chr1.

1. Converting wiggle file into text file using custom script reformatWig.pl as R is not memory efficient in data frames with millions of rows.

   ```
   perl reformatWig.pl –i chr1.wig –o chr1.txt
   ```

   The output file chr1.txt is a BED file with start, end positions and the coverage of this range.

2. Peak calling using peak.calling.R as command line

```
R CMD BATCH --no-save --no-restore '--args
data.file="chr1.txt" out.file="chr1.peaks.txt" pctile=95'
peak.calling.R peak.calling.chr1.out
```

The output file `chr1.peaks.txt` is a tab delimited text file including the peak position, peak height, peak start and peak end positions.

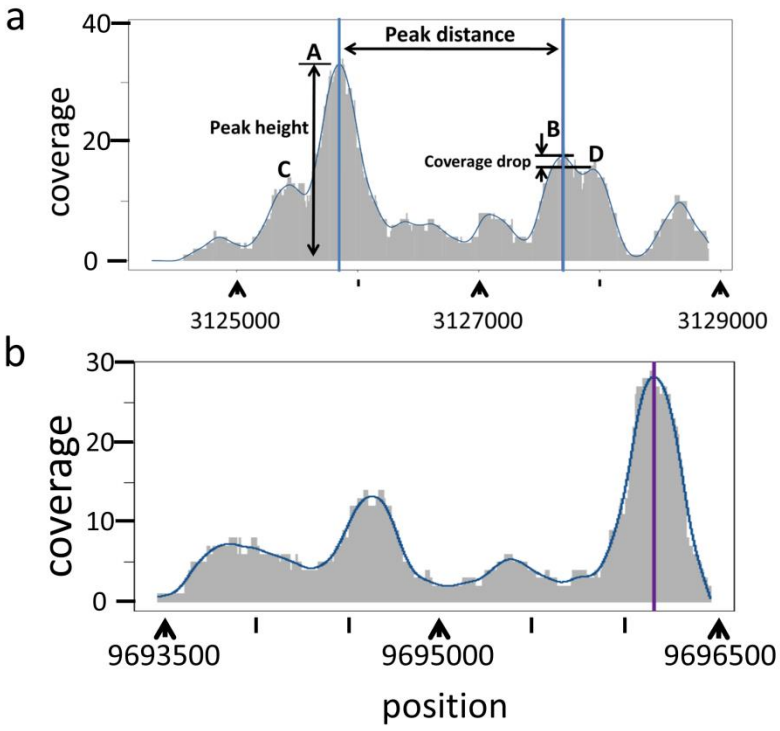

Figure S4. Examples of replication origin identification and motif search results.
a) Examples of identified replication origins (A and B) using coverage data. The grey shading shows coverage across the genomic region, the blue curve is the smoothed coverage, and the purple lines show the detected peaks as potential replication origins. The height of peak C is less than the 95% cutoff value (with coverage ~14) and is excluded as a replication origin. Peak A, B, and D are all have coverage greater than the cutoff coverage values (14 in this case), peak D does not have enough coverage drop between peak C and is not considered as an origin either. b) Example of a peak detected by the smooth.spline R function used here but not detected by MACS due to the wider genomic range and overlapping peaks (the missing peak is indicated with the purple line).

F. Coverage normalization and difference wiggle files generation between experimental and control samples

The coverage wiggle files are normalized based on total number of mapped reads using a 10bp window based approach in control and experimental samples with sex

chromosomes excluded to account for the potential gender differences. For example, the counts of total mapped reads in wild type control and experimental Mcm2 deficient samples for the first experiment [7] are 323336800 and 281295797 respectively, the normalization scaling factor would be 1.15 (323336800/281295797 = 1.15). The coverage value for experimental sample will be normalized by multiplying the scaling factor. After normalization, the difference wiggle files are generated by simply subtracting control sample coverage wiggle file from experimental coverage wiggle file.

CAUTION: It is important to perform a correction for sex chromosomes as your control and experimental samples may come from different sex individuals.

Commands:

1. get total coverage from wiggle files:

```
cut -f 2 *.wig |grep -E "^[0-9]" | awk '{s+=$1} END \
{print s}'
```

2. make wiggle difference file

```
java -jar WigSubtractor.jar wigglefile_1 wigglefile_2
```

G. Sequence feature search for replication origins

BEDtools [8] is used to identify unique peaks for control and experimental samples.

```
bedtools intersect -a experimental.peaks.bed \
     -b control.peaks.bed \
     -v > xperimental.uniq.peaks.bed
```

DNA sequences around previously identified peaks are extracted using Picard-tools. For each peak, 2000 bases before and after the peak position are extracted for sequence analyses including known motif identification, new motif search, and stiffness/bendability calculation.

```
java -jar RetrievePeakSequence.jar ref.fa input_peaks.bed \
output.fa
```

H. G-quadruplex and TG repeat search

These sequence motifs have been reported to be present at replication origins. The exact position distribution around replication origins and the proportion of origins with such motifs are still unknown. A simple regular expression search in the sequences around peaks is carried out to identify the positions of all G-quadruplexes and TG repeats (4or greater). The regular expressions used for these two sequences are G{3,}.{1,7}G{3,}.{1,7}G{3,}.{1,7}G{3,} and (TG){4,} respectively. Then the frequencies of these two motifs at each base position are summarized from the motif search and plotted as histogram.

Commands:

```
perl motifSearch.pl -s '(G{3,}).{1,7}\1.{1,7}\1.{1,7}\1' \
     -i  sample.peaks.fa --verbose > gquards3gs.txt
```

```
perl motifSearch.pl -s '(TG){4,}' -i sample.peaks.fa \
     --verbose > TG4.txt

grep -h -v '#' gquads3gs.txt | cut -f 3,4,5 | perl -e \
     'while(<>){
          chomp;@x=split(/\t/,$_,-1);
           if($x[2] eq "-")
          {
                 $x[0]=4002-$x[0];$x[1]=4002-$x[1];
           }
           $mid=int(($x[1]+$x[0])/2);
           $hash-> {$mid}++;
     }
     foreach my $p (1..4001){
           if(exists($hash->{$p})){
                 print("$p\t$hash->{$p}\n")
           }
     }' > gquads3gs.txt.hist


grep -h -v '#' gquadsTG4.txt | cut -f 3,4,5 | perl -e \
     'while(<>){
           chomp;
           @x=split(/\t/,$_,-1);
           if($x[2] eq "-"){
                 $x[0]=4002-$x[0];
                 $x[1]=4002-x[1];
           }
           $mid=int(($x[1]+$x[0])/2);
           $hash->{$mid}++
     }
     foreach my $p (1..4001){
           if(exists($hash->{$p})){
                 print("$p\t$hash->{$p}\n")
           }
     }' > TG4.txt.hist
```

I. Novel motif search

To identify potential sequence motifs of replication origins, we carried out motif discovery and search using MEME software suites [9]. To identify sample specific motifs, DREME [10] program, which identifies enriched sequence motifs in the test sequence against background sequence, from the MEME suites is used by treating another sample as background. During motif search, the length of motifs is restricted between 5 and 15 bases, and the motif could appear one or more times in each origin. Due to the large number of putative replication origins identified in the analysis, it is unfeasible to run MEME/DREME on all sequences. Alternatively, motif discoveries are done on each chromosome separately. For each identified motif, MAST [11] motif searches are used across the genome to identify the occurrences. The histogram for each motif is also plotted.

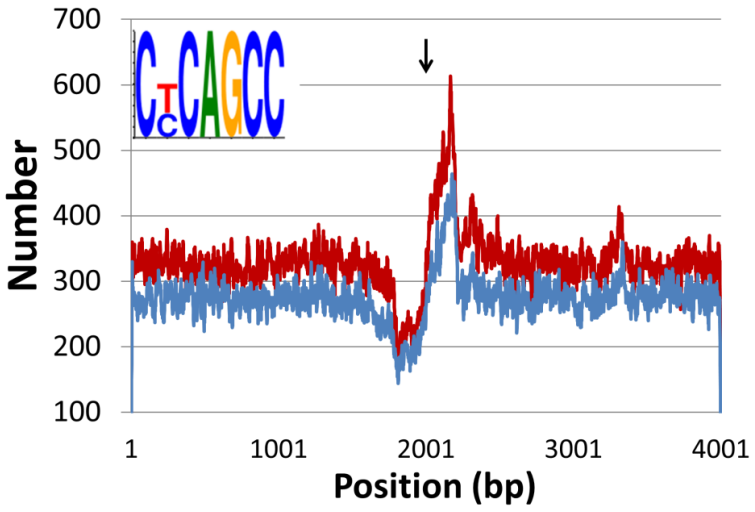

Figure S5. Motif identified by DREME.
Example of a GC-rich motif (inset, upper left) identified by DREME and histograms showing the number of occurrences of the motif at each position within 2kb of peak maxima (arrow) from wild type (blue, out of 226665 peaks) and Mcm2 deficient (red, out of 225073 peaks) samples. The motif shows an asymmetric distribution relative to the peak maxima. Further, it is enriched in the vicinity of peaks from Mcm2 deficient, relative to wt, cells.

Commands of running DREME for each chromosome using shell script:

```
for i in {1..19}, X, Y
do
      dreme -o treat-ctrl -mink 5 -maxk 15 -p \
      treat.chr$i.peak.fa -n ctrl.chr$i.peak.fa
done
```

To run MAST, go to their website (http://meme-suite.org/tools/mast) to upload peak files and the motifs identified in this step.

J. Stiffness and bendability analysis

Other than sequence motifs, replication origin could be determined by high level structure of the DNA. Stiffness and bendability are two different measures for calculating DNA curvature. For each sequence around the identified potential origin, the stiffness and bendability along the sequence are calculated using the tri-nucleotide bendability from the bend-it server [12] and stiffness scale [13]. The position specific distribution of stiffness and bendability for all putative origins are plotted. For sequence motifs discovered in previous steps, the stiffness and bendability curves are also generated.

Commands:

```
perl calculateStiffness.pl output_stiff.txt input.peak.fa \
      stiff_coef.txt
perl calculateBendability.pl output_bend.txt input.peak.fa\
      bend_coef.txt
```

VI. Supplementary References:

1. Frick DN, Richardson CC (2001) DNA primases. Annu Rev Biochem 70: 39-80.
2. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.
3. Li H et al. Genome Project Data Processing S: (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
4. Zhang Y et al, (2008) Model-based analysis of ChIP-Seq (MACS). Genome biology 9: R137.
5. Kent WJ et al. (2002) The human genome browser at UCSC. Genome research 12: 996-1006.
6. Heinz S, at al. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell 38: 576-589.
7. Kunnev D et al. (2015) Effect of minichromosome maintenance protein 2 deficiency on the locations of DNA replication origins. Genome Res 25: 558-569.
8. Quinlan AR & Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.
9. Bailey TL, et al. (2009) MEME SUITE: tools for motif discovery and searching. Nucleic acids research 37(Web Server issue): W202-208.
10. Bailey TL (2011) DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics 27: 1653-1659.
11. Bailey TL & Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. Bioinformatics 14: 48-54.
12. Vlahovicek K, Kajan L, Pongor S (2003) DNA analysis servers: plot.it, bend.it, model.it and IS. Nucleic acids research 31: 3686-3687.
13. Gromiha MM (2000) Structure based sequence dependent stiffness scale for trinucleotides: a direct method. Journal of biological physics 26: 43-50.