The location and development of Replicon Cluster Domains in early replicating DNA

Melo da Costa Nunes, Jose; Gierlinski, Marek; Sasaki, Takayo; Haagensen, Emma J.; Gilbert, David M.; Blow, J. Julian

Published in:
Wellcome Open Research

DOI:
10.12688/wellcomeopenres.18742.2

Publication date:
2023

RESEARCH ARTICLE

REVISED The location and development of Replicon Cluster Domains in early replicating DNA [version 2; peer review: 3 approved]

José A. da Costa-Nunes[1], Marek Gierlinski[2], Takayo Sasaki[3], Emma J. Haagensen[1,4], David M. Gilbert[3], J. Julian Blow[1]

[1]Division of Molecular, Cell and Developmental Biology, School of Life Sciences, University of Dundee, Dundee, DD1 5EH, UK
[2]Data Analysis Group, School of Life Sciences, University of Dundee, Dundee, DD1 5EH, UK
[3]San Diego Biomedical Research Institute, San Diego, California, CA 92121, USA
[4]Present address: School of Medical Education, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, NE2 4HH, UK

v2 **First published:** 11 Apr 2023, **8**:158
https://doi.org/10.12688/wellcomeopenres.18742.1
Latest published: 22 Aug 2023, **8**:158
https://doi.org/10.12688/wellcomeopenres.18742.2

Abstract

Background: It has been known for many years that in metazoan cells, replication origins are organised into clusters where origins within each cluster fire near-synchronously. Despite clusters being a fundamental organising principle of metazoan DNA replication, the genomic location of origin clusters has not been documented.
Methods: We synchronised human U2OS by thymidine block and release followed by L-mimosine block and release to create a population of cells progressing into S phase with a high degree of synchrony. At different times after release into S phase, cells were pulsed with EdU; the EdU-labelled DNA was then pulled down, sequenced and mapped onto the human genome.
Results: The early replicating DNA showed features at a range of scales. Wavelet analysis showed that the major feature of the early replicating DNA was at a size of 500 kb, consistent with clusters of replication origins. Over the first two hours of S phase, these Replicon Cluster Domains broadened in width, consistent with their being enlarged by the progression of replication forks at their outer boundaries. The total replication signal associated with each Replicon Cluster Domain varied considerably, and this variation was reproducible and conserved over time. We provide evidence that this variability in replication signal was at least in part caused by Replicon Cluster Domains being activated at different times in different cells in the population. We also provide evidence that adjacent clusters had a statistical preference for being activated in sequence across a group, consistent with the 'domino' model of replication focus activation order observed by microscopy.
Conclusions: We show that early replicating DNA is organised into

Open Peer Review

Approval Status ✓ ✓ ✓

	1	2	3
version 2 (revision) 22 Aug 2023	✓	✓ view	✓ view
version 1 11 Apr 2023	? view	? view	? view

1. **Ichiro Hiratani**, RIKEN Center for Biosystems Dynamics Research, Kobe, Japan

2. **Itamar Simon**, Hebrew University of Jerusalem, Jerusalem, Israel
Avraham Greenberg, Hebrew University of Jerusalem, Jerusalem, Israel

3. **Zhongqing Ren**, Indiana University Bloomington, Bloomington, USA

Any reports and responses or comments on the article can be found at the end of the article.

Replicon Cluster Domains that behave as expected of replicon clusters observed by DNA fibre analysis. The coordinated activation of different Replicon Cluster Domains can generate the replication timing programme by which the genome is duplicated.

Keywords
DNA replication, S phase, replicon clusters, replication timing, cell cycle

Corresponding author: J. Julian Blow (j.j.blow@dundee.ac.uk)

Author roles: da Costa-Nunes JA: Investigation, Methodology, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Gierlinski M**: Data Curation, Formal Analysis, Software, Writing – Review & Editing; **Sasaki T**: Investigation, Methodology, Writing – Review & Editing; **Haagensen EJ**: Investigation, Methodology, Writing – Review & Editing; **Gilbert DM**: Conceptualization, Funding Acquisition, Supervision, Writing – Review & Editing; **Blow JJ**: Conceptualization, Formal Analysis, Funding Acquisition, Project Administration, Software, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome [096598, https://doi.org/10.35802/096598].

How to cite this article: da Costa-Nunes JA, Gierlinski M, Sasaki T *et al.* **The location and development of Replicon Cluster Domains in early replicating DNA [version 2; peer review: 3 approved]** Wellcome Open Research 2023, **8**:158 https://doi.org/10.12688/wellcomeopenres.18742.2

First published: 11 Apr 2023, **8**:158 https://doi.org/10.12688/wellcomeopenres.18742.1

REVISED **Amendments from Version 1**

In response to the reviewers' comments we have:

Re-written some text to provide additional clarification.

Added Repli-Seq (timing domain) data to Figure 8.

Added a heatmap colour scheme to Figure 2C.

Added statistical test data to Figure 7, Figure 9 and Figure 10.

Re-ordered the permutations in 10A and B so that in moving left to right it moves from less domino-like to more domino-like.

Replotted the FACS data in Figure 1 and Supplementary Figure S1 to decrease the extent of the y axis (DNA content) thereby making the increase in DNA content more obvious.

Corrected various minor typos and errors.

Any further responses from the reviewers can be found at the end of the article

Introduction

The human genome harbours approximately 6.6 Gbp of DNA that is split between 46 chromosomes[1] and is packed in the nucleus in chromatin bearing different degrees of compaction[2]. Almost all of the DNA is replicated during S phase of the cell division cycle. In somatic human cells, S phase typically lasts around 10 hours, and the full cell cycle lasts approximately 24 hours[3,4].

In late mitosis and G1, prior to the initiation of DNA replication in S phase, the sites of where replication initiation can take place (DNA replication origins) are licensed by being encircled by double hexamers of the MCM2-7 proteins. However, more DNA replication origins are licensed in G1 than are activated during S phase[5]. Indeed, of the total number of potentially available replication origins present in diploid human cells – roughly 500,000 – only approximately 10% will be used to initiate DNA replication[3,5–7]. The segment of DNA replicated by the advancement of the fork or forks emanating from one origin is called a replicon and has an average size of ~150 Kbp[4,7–9]. Given a fork speed of 1–2 kb/min[8,10,11] an average replicon is therefore active for 37.5–75 min. Some 'strong' replication origins are active in almost all cells in a population, but for weaker origins there is a high degree of cell-to-cell variability in exactly which origins fire and which remain dormant[7,12–14].

Extensive data from DNA fibre analysis has shown that replication origins are typically organised into clusters of two to ten origins (typically from three to six) that fire near-synchronously during S phase[8,15]. The average size of an origin cluster would therefore be in the range of 400–800 kb. The clustering of initiation events at particular loci may be a consequence of origins with similar replication timing being grouped together in regions[7]. A recent paper has provided evidence that replication initiation domains are characterized by highly efficient origins flanking a cluster of less efficient origins[14]. However, to date, there is little understanding of the actual DNA sequences that make up origin clusters.

At a higher level, it is observed that large (megabase) regions of DNA replicate at characteristic times in S phase; these regions are termed 'replication timing domains' (RTDs)[16–21]. With the advent of next generation sequencing, it became possible to identify the different timing domains and correlate them with higher-order nuclear architecture[22–24]. Each cell line or cell type has a characteristic timing domain profile which can then change if the cells differentiate[22,25–27]. Timing domains vary in size, ranging from a few hundred Kbp to around 10 Mbp[22,28,29]. Their sizes suggest that timing domains typically comprise more than one adjacent replicon cluster[22,30]. The conservation of the replication timing profiles is in contrast to the considerable stochasticity in the firing of the individual origins within each timing domain.

At the cytological level, DNA replication occurs in replication *foci* within the nuclei of S phase cells. The subnuclear localisation of these replication foci changes in a predictable way during the course of S phase[8,15,18,31–33], reflecting progression through the replication timing programme. There are ~1,000 *foci* active at any given time in a typical somatic S phase cell[34], and with ~10,000 replication forks being active at any one time in S phase, each replication *foci* contains ~10 active DNA replication forks[34,35]. The similarity between the size and distribution of origin clusters (as observed by DNA fibre analysis) and replication *foci* (as observed by microscopy) suggests that they probably represent the same fundamental unit of DNA replication[8,15,30,36]. The DNA replicated in individual replication foci persist as foci throughout G1, S, G2, and mitosis over multiple cell cycles, suggesting that they represent a stable unit of chromosome organisation[8,15].

In this paper we sought to identify genomic regions that correspond to the replicon clusters observed by DNA fibre analysis. We use a synchronisation protocol that gives a cohort of cells that enter S phase with a high degree of synchrony. At different times after release into S phase, we pulsed cells with EdU and sequenced the DNA replication ongoing during the length of time that these pulses lasted. Our high-resolution data shows the existence of discrete peaks of DNA synthesis within individual Early Timing Domains. The width of these peaks (~500 kb) is consistent with them representing individual replicon clusters, so we have named them 'Replicon Cluster Domains'. The distribution and evolution of these Replicon Cluster Domains (RCD) suggest that their activation time varies between different cells in a population and shows a preference for a sequential 'domino'-type activation sequence.

Methods

Cell culture

U2OS human osteosarcoma bone cells (U-2 OS, ATCC® HTB-96) were purchased from ATCC. Neuro-2a mouse neuroblasts cells (ATCC® CCL-131™) were also used as a control in some experiments. Cells were grown in Tissue Culture incubators (Forma Scientific CO2, water, jacketed

incubator; Life Sciences Instruments), in pre-warmed (37°C) 1xDMEM (containing 4,500 mg/L glucose, 110 mg/L sodium pyruvate, 584 mg/L L-glutamine and no HEPES) (Dulbecco's Modified Eagle Medium) (Thermo Scientific/Gibco, 41966-052) medium complemented with FBS (Thermo Fisher Scientific/Gibco, 10270-106) (10% v/v/), and 100 U/ml of penicillin and streptomycin (P/S) ((Fisher Scientific/Gibco, 15140-122) (1% v/v/), at 37°C, 5% CO_2. Cells washes were carried out with 1xDPBS (Thermo Scientific/Gibco, 14190-169) pre-warmed at 37°C.

Cells were harvested, before reaching confluency, after Trypsin-EDTA (0.05%) (Gibco, 25300-054) digest at 37°C, for 5 min; digest was stopped by the addition of 1xDMEM (pre-warmed at 37°C). Cells were spun in a centrifuge (Eppendorf centrifuge 5810) at 211xG (1.000 rpm), for 5 min at room temperature (RT), and the supernatant removed. The cell pellet was either immediately frozen in liquid nitrogen or was resuspended in a small volume of 1xDMEM and fixed in ice cold (-20°C) 70% ethanol solution overnight (or for several days) at -20°C.

Cell cycle synchrony (G0/G1 cell cycle arrest)

Cells were first grown in pre-warmed (37°C) 1xDMEM (Dulbecco's Modified Eagle Medium) medium complemented with FBS (10% v/v/) and penicillin and streptomycin (P/S) (1% v/v/), at 37°C, 5% CO_2. After 24 hours the medium was replaced with fresh medium 1xDMEM complemented with 0.1% (v/v/) FBS and 1% (v/v) P/S, and cells were grown for three more days. Cells washes were carried out with 1xDPBS pre-warmed at 37°C. Cells were washed with 1xDPBS prior to trypsin digest, and cell harvest was carried out as described in Cell cycle synchrony (G1/S phase cell cycle arrest).

Cell cycle synchrony (G1/S phase cell cycle arrest)

Cells were inoculated in 1xDMEM+FBS+P/S medium and allowed to grow for 30 hours. Cells were then grown in fresh 1xDMEM+FBS+P/S medium containing Thymidine (2.5mM) (Sigma, T1895-5g) for a further 24 hours. Cells were then grown for 10 hours in 1xDMEM+FBS+P/S medium only. Finally, the cells were grown for six hours in 1xDMEM+FBS+P/S medium containing L-mimosine (0.5mM) (Sigma, M0253-100mg). All medium was pre-warmed at 37°C prior to being added to the cells.

After the L-mimosine treatment, cells were immediately washed with ice cold (4°C) 1xPBS, prior to adding fresh medium (pre-warmed at 37°C) to the cells. Cells were harvested using Trypsin-EDTA (0.05%) (5 min incubation, at 37°C), followed by enzymatic neutralisation by adding cold (4°C) 1xDMEM. Cells were immediately collected and spun in an Eppendorf centrifuge (5804R) (at 4°C) for 10 minutes, at 180 rcf. The cell pellet was either immediately frozen in liquid nitrogen or was resuspended in a small volume of 1xDMEM and fixed in ice cold (-20°C) 70% ethanol solution overnight (or for several days) at -20°C. These cell cycle synchronised cell populations are referred to, in this paper, as TM cell cycle cell populations.

Incubation with EdU

Prior to harvesting cells were incubated with 1xDMEM+FBS+P/S medium containing 40 μM 5-ethynyl-2'-deoxyuridine (EdU) (Invitrogen, A10044) for different amounts of time, depending on the time course experiment being carried out. EdU was used from a stock of 10 mM EdU in dimethyl sulfoxide (DMSO); control cells not exposed to EdU had the same amount of DMSO added to the medium.

Cell culture (for next generation sequencing - NGS)

Approximately one million cells were inoculated in 20 ml medium in a 150 mm diameter plate. Each biological replica, for each time point experiment, was carried out using three 150 mm diameter plates. Cell cycle synchronous cells experiments were prepared as described above. Asynchronous cells experiments were carried out as described above (in Cell culture). After digestion with Trypsin-EDTA and neutralisation with 1xDMEM (at 4°C), a small aliquot of the re-suspended cells was collected, and spun in an Eppendorf 5804R centrifuge, at 180 rcf, for 5 min (at RT) and fixed in 70% ethanol solution at -20°C; these cells were used to do cell cycle analysis by flow cytometry. The remaining re-suspended cells were centrifuged at 4°C, for 10 min, at 180 rcf; the cell pellet was immediately frozen in liquid nitrogen.

2D flow cytometry analysis (2D FACS)

The Click-iT reaction (Click-iT Plus EdU Alexa Fluor 647 (flow cytometry assay kit); Invitrogen, C10635) was carried out on cells fixed in 2 ml of 70% ethanol (at -20°C) for at least 12 hours, in the dark. To the 2 ml of cells fixed in 70% ethanol at -20°C O/N, 3 ml of filtered 1% BSA (in 1xPBS) were added, followed by centrifugation in an Eppendorf centrifuge (5804R) at 260xG for 5 min at 4°C. The 1% (w/v) bovine serum albumin (BSA) fraction V (Roche, 10735094001) (in 1xPBS) solution was filtered with a 0.22 μm filter (Fisher, 10400031) prior to use. After removal of the supernatant, the pellet was resuspended in approximately 100 μl of remnant supernatant, and washed one more time with 1 ml of 1% BSA (in 1xPBS) followed by a centrifugation at 260xG for 5 min at 4°C. After removal of the supernatant, 250 μl of the Click-iT reaction mix (2mM $CuSO_4$ II, 0.1x reaction buffer additive, 1:200 dil. Alexa Fluor 647, in 1xPBS) (Click-iT Plus EdU Alexa Fluor 647 flow cytometry assay kit; Invitrogen, C10635) was added to the cells. The Click-iT reaction was carried out in the dark, at room temperature (RT), for 30 min. The reaction was terminated by adding 0.8 ml 1% BSA (in 1xPBS). After a centrifugation at 260xG for 5 min at 4°C, and removal of the supernatant, 0.8 ml of 1% BSA (in 1xPBS) were added to the cell pellet. Cells were centrifuged again and resuspended in 500 μl propidium iodide (50 μg/ml propidium iodide (Sigma, P4864-10ml), 50 μg/ml RNaseA (DNase-free, protease-free) (Thermo Scientific, EN0531), in 1xPBS). Cells were kept in the dark for 30 min at RT, prior to flow cytometry.

Samples were analysed on a FACS Canto (Becton Dickinson) flow cytometer, using FACSDiva version 8.01. Propidium iodide was detected using 488 nm excitation and emission was detected at 530/30 nm. AF647 fluorescence was detected

using 640 nm excitation and emission detected at 660/20 nm. Data analysis was carried out using FlowJo software, version 10.6.2. Whilst FlowJo is one of the leading software packages available for analysing flow cytometry data, the data can also be analysed by free software packages such as Cytospec.

Cell sorting

Cells were sorted on the Influx (Becton Dickinson) or on the SH800 (Sony Biosciences). Results yielded from both cell sorters did not differ significantly. In both cases, propidium iodide fluorescence was detected using 488nm excitation and fluorescence detected at 580/30nm on the Influx and 600/60nm on the SH800.

All samples were cell sorted except the asynchronous (AS) cells labelled with EdU for one hour and for 24 hours (AS1 and AS24); the entirety of the cells harvested from the cell culture plates from these samples (AS1 and AS24) was used in the production of the respective gDNA libraries. The quality of these samples was checked by 2D FACS based on propidium iodide (PI) (Sigma, P4864-10ml) fluorescence signal and Alexa Fluor 647 (Click-iT reaction) fluorescence signal. All the other cell samples that were cell sorted, were first PI stained, and RNAseA treated (Thermo Scientific; EN0531), prior to being sorted/collected.

Genomic DNA extraction

Genomic DNA (gDNA) extraction was carried using the DNeasy Tissue & Blood kit (Qiagen, 69504), following the protocol recommended (for mammalian cells) by the manufacturer. Elution of the genomic DNA from the columns was carried out in two consecutive elution steps each with 200 µl 1xT.E. (0.1mM EDTA) pH 8.0. The integrity of the extracted gDNA was checked in an agarose (0.8% (w/v) agarose in 1xTBE) electrophoresis gel, and was quantified in a spectrophotometer (Geneflow Nanophotometer, Geneflow).

H_2O was added to the eluted gDNA to give a final volume of 250 µl. The gDNA was sonicated in an ice-cold water bath in a Diagenode Bioruptor (high power, 15 min) with alternating 30 second cycles of sonication. The size range of gDNA fragments checked on a 2.5% agarose gel ranged from ~100–900bp, with the bulk being between ~200–600bp. Sonicated gDNA was precipitated (0.31 µg/µl glycogen, 70% ethanol (molecular biology grade), 83mM NaOAc pH 5.2), overnight at -80°C. After centrifugation (20817xG, 40 min, 4°C) the pellet was washed twice with 70% ethanol, dried and resuspended in 15 µl 10mM Tris-HCl pH 7.4.

Click-iT reaction on sonicated gDNA

Click-iT Nascent RNA Capture kit (Life Technology/Thermo Fisher, C10365) was used to biotinylate the EdU present in the sonicated gDNA, following the manufacturer's instructions. The 50 µl Click-iT reactions were precipitated (50 µg/ml glycogen, 0.47M NH_4OAc, 87.5% ethanol, -80°C overnight); after centrifugation (20817xG, 40 min, 4°C) pellets were washed twice with 75% ethanol, dried and resuspended in 40 µl H_2O. The samples were quantified in a spectrophotometer (Geneflow Nanophotometer, Geneflow).

gDNA library construction

For each sample, three reactions of End-repair (with 1 µg of gDNA per reaction) were carried out using the Next ULTRA library prep. kit for Illumina (NEB, E7370S). This was followed by the ligation of the Next Adaptor for Illumina to the ends of the DNA fragments (NEB Next Multiplex Oligo for Illumina (index primer 1) kit (NEB, E7335L)). Finally, USER enzyme digest was carried out. The manufacturer's protocol was followed in all these three steps. After the digest with the USER enzyme (NEB, M5505S), the DNA was precipitated (0.47M NH_4OAc, 50 µg/ml glycogen, 87.5% ethanol, -80°C, overnight) and resuspended in H_2O. The three reactions from each sample were bulked together. DNA was purified using a (Mini elute PCR purification kit (Qiagen, 28004) and the eluted DNA quantified in a spectrophotometer. One percent of the volume of the eluted samples was used for sequencing the full-length genome DNA of that same sample (the 1% sequence data). The remaining 99% of the samples were used for the biotin pull-down.

Biotin-tagged gDNA library pull-down for next generation sequencing

Pull-down was carried out with all gDNA libraries except the gDNA libraries made from the G0/G1 cells. Pull-down was carried out using streptavidin coated magnetic Dynabeads (MyOne C1, Thermo Fisher/Invitrogen, 65001) and low adherence tubes (Axygen max. recovery, 1.5 ml, Corning, MCT-150-L-C). In each pull-down, 30 µl magnetic beads were used. Beads were washed three times with cold 1x B&W buffer (5mM Tris-HCl (pH 7.4), 0.5mM EDTA (pH 8.0), 1M NaCl); 60 µl cold 2x B&W buffer was added to the beads plus the biotinylated-gDNA. Binding was carried out in the dark, for >1 hour at room temperature on a tube roller (roller mixer SRT9, Stuart). Tubes were then placed in a magnetic rack (Genetics/FastGene MagnaStand 1.5, FG-SSMAG1.5) for 3 min, and all solution was removed. Beads were washed with cold 1x B&W buffer; the tube was then placed in the magnetic rack for 3 min and the supernatant was removed. After three more washes the beads were washed with H_2O and then suspended in 25 µl H_2O to give the pulled-down sample. For quality control of the pull-down, 1 µl of this sample was used; another 1 µl was used for qPCR to determine the appropriate indexing/amplification cycle. The remaining 23 µl of the pulled-down sample was used for the indexing of the library, using a combination of universal primer and an index primer (NEB Next Multiplex Oligo for Illumina (index primer 1; NEB, E7335L). The indexing PCR reaction pulled-down sample was carried out directly on the beads, using a 1:100 dilution of the bead's suspension and primers for the adaptor region (NEBadqPCR_F; ACACTCTTTCCCTACACGACGC and NEBadqPCR_R; GACTGGAGTTCAGACGTGTGC) then dual-indexing was done using NEB E7600S or E7780S, aiming to obtain a total of 100–1,000 ng indexed DNA (typically 8–18 cycles of amplification).

Quality control of the pull-down, prior to the indexing of the library, was carried out in Eppendorf Mastercycler PCR machine as follows: First, the 1 µl of beads from the pulled-down sample and a small volume of the input 1% was amplified

with eight PCR cycles using the Illumina library kit (NEB Q5 Hot Start PCR Master mix) and Next ULTRA library prep. kit for Illumina (NEB, E7370S) plus the Universal primer and the index primer 1 (NEB Next Multiplex Oligo for Illumina; NEB, E7335L) in 20 µl PCR mix (1 µM of Index 1 primer, 1 µM Universal primer, 0.5x NEB Q5 Hot Start buffer + enzyme), 1x cycle at 98°C for 30 sec, 8x cycles at 98°C for 10 sec, followed by 65°C for 1 min 15 sec, and a final 1x cycle at 65°C for 5 min. A second PCR amplification of 45 cycles was carried out on 4 µl of the product of the first PCR amplification without beads, using the same primers but another Taq enzyme and PCR buffer (Thermo Fisher/Invitrogen, 10342053) in 30 µl of a PCR mix (1x PCR buffer Thermo Taq, 0.25mM dNTP, 1 µM of Index 1 primer, 1 µM Universal primer, 3 units Thermo Taq). This second PCR reaction had the following parameters: 1x cycle at 95°C for 2 min, 45x cycles at 95°C for 20 sec, 65°C for 30 sec, 68°C for 1 min 15 sec, and a final 1x cycle at 68°C for 5 min. The purpose of this second PCR was to determine whether the PCR products (run in a 2.5% agarose gel in 1xTAE buffer) matched the gDNA smear observed when the gDNA library itself was run on a 2.5% agarose gel. In the third PCR, 0.5 µl of the products of the first PCR amplification (minus beads) was used with h_intGenMit-F ($^{5'}$CCTAGGAATCACCTCCCATTCC$^{3'}$) and h_intGenMit-R ($^{5'}$GTGTTTAAGGGGTTGGCTAGGG$^{3'}$) primers[37], as well as Taq enzyme and PCR buffer (Thermo Fisher/Invitrogen, 10342053) in 20 µl PCR mix (1x PCR buffer Thermo Taq, 0.25mM dNTP, 0.8 µM of h_IntGenMit-F primer, 0.8 µM h_IntGenMit-R primer, 1 unit Thermo Taq). The cycling parameters of the third PCR reaction were: 1x cycle at 95°C for 2 min, 33x cycles at 95°C for 45 sec, 60°C for 1 min, 68°C for 2 min, and a final 1x cycle at 68°C for 5 min. The purpose of this third PCR was to assess if the pull-down samples contained pulled-down DNA from the U2OS cells.

Sequencing data processing

Quality control of FASTQ reads was carried out with *FastQC* ver. 0.11.4 and revealed significant contamination with adapter sequencies. Subsequently, the adapters were trimmed using *TrimGalore* ver. 0.5.0. Reads were mapped to the human reference genome GRCh38, release 94, obtained from Ensembl, using *Bowtie2* ver. 2.3.0[38]. Resulting BAM files were filtered for read quality (MAPQ > 10), sorted and indexed with *Samtools* ver. 1.9[39]. BAM files were converted into BED files, which were binned into bedGraph format in 10,000- and 50,000-bp bins, using *Bamtools* ver. 2.27.1[40]. The scripts created for this analysis can be found in the Github repository:

https://github.com/bartongroup[41]

Background subtraction and normalisation

bedGraph files contain distribution of reads (genomic tracks) for both pull-down (P) and genomic input control (C). Background subtraction and normalisation is based on the CISGenome normalization[42,43]. For a given pull-down and control the score is defined as reads per million, P_i and C_i, respectively, where i indicates position in the genome (across all chromosomes). We assume that the pull-down consists in part of the genomic background, B_i, since pull-down procedure is never 100% effective and the real pull-down signal: $P_i = B_i + S_i$. We assume that the background present in the pull-down has the same distribution across chromosomes as the control, with an unknown normalisation factor, $B_i = rC_i$. The purpose of normalisation and background subtraction is to find r.

A plot of log (P_i/C_i) versus log $(P_i + C_i)$ can be approximated by a broken line with two segments. The horizontal segment, at low counts, represents genomic regions were pull-down and background are equal, that is with no signal, $S_i = 0$. A change in total count, $P_i + C_i \approx (1 + r)C_i$ does not affect a constant ratio of $P_i/C_i \approx r$. In contrast, the part of the plot with a positive slope represents regions with positive signal $S_i > 0$ where $P_i/C_i = r + S_i/C_i$. Thus, $P_i/C_i \propto S_i$ where signal is strong, $S_i \gg B_i$. By fitting these data with a broken line, a break point b can be found. All data with log $(P_i + C_i) < b$ belong to genomic regions with no signal. This set is denoted as $G = \{i: \log(P_i + C_i) < b\}$. The CISGenome method finds the normalisation factor as

$$\hat{r}_{cis} = \frac{\sum_{i \in G} P_i}{\sum_{i \in G} C_i}.$$

This approach, however, does not work well with data with high coverage of peaks, as the background-isolation method is not perfect and G contains, in part, regions with some signal. The distribution of P_j/C_j, where $j \in G$, is not symmetric and contains a high-count tail. Here, the peak of the P_j/C_j distribution (the mode) is chosen to estimate the normalisation factor $\hat{r}$, as it represents the most frequently found P_j/C_j ratio. After this, the background-subtracted signal is found as $S_i = P_i - \hat{r}C_i$, for each bin i.

Data analysis and deposition

Preliminary data investigation and analysis, normalisation and background subtraction were done in R. The code is available at GitHub (https://github.com/bartongroup) and archived with DOI 10.5281/zenodo.7639072[41]. The 10–40, 40–70, 70–100 and 100–130-minute data used in most of the paper came from a single experiment that was processed in parallel, but the 0–40 min data came from a separate experiment. All datasets are provided at Biostudies (S-BSST966). The replication timing data for U2OS can be derived from data provided at https://data.4dnucleome.org/, accession numbers: 4DNES99LXRYK and 4DNES1P18J2X. The final data as used in this study are available at the Biostudies (S-BSST966 site, RT_U2OS_Bone.txt)

Further analysis was done in the Swift programming language (Swift 5.7.2) using the Xcode development environment (Xcode version 13.3). The analysis code is available together with the R code at GitHub[41]. It consists of: i) a DataCentre class which contains all the core analysis functions; ii) an EarlyRepDataSet class which stores the early replication signal and derived information from it, including the wavelet analysis signals; iii) a U2OSTimingDomains class which stores the information from the previously published replication timing analysis on U2OS cells; iv) a WaveletAnalysis file

which contains the functions for performing wavelet analysis with the Ricker wave; v) an adjacentValueSimilarityMetric function that returns a metric indicating the similarity between adjacent values in an input array; vi) a Gaussian struct which supports the Gaussian function; vii) a flatToppedGaussian struct that supports a Gaussian function with a flat middle section; viii) a NelderMead class that performs parameter fitting using the Nelder-Mead algorithm; ix) an AppDelegate class which provides Controller function to mediate between the graphical user interface and the analysis classes; and x) View structures and classes that provide the graphical user interface (MainMenu, JBGraph, Decimal+Normalisation and JBGraphSheets). The operation of most of the functions are described in the main text and figure legends.

The adjacentValueSimilarityMetric function works as follows. Given an input array of n values, the function returns a metric which indicates the similarity between adjacent values in the input array. These values are based on the n-1 absolute differences between adjacent values in the array. First, the function calculates the mean value of the absolute differences between adjacent values in the array (AdjacencyDifferenceForArray; ADFA). The function then calculates the mean value of the n-1 absolute differences between adjacent values in a sorted version of the array (AdjacencyDifferenceForSortedValues; ADFSV); this is equivalent to the absolute difference between the largest and smallest values divided by n-1. Next, the function calculates the mean value of the absolute differences between adjacent values in randomly-sorted versions of the array (MeanAdjacencyDifferenceForAllPermutations; MADFAP); this is equivalent to the average difference between all non-identical pairwise comparisons of the values in the array. The AdjacencyDifferenceForArray metric is then normalised so that a value of one means that the adjacent values are ordered in a maximally similar way; a value of zero means random similarity; and negative values indicate anti-similarity. This is calculated for the adjacentValueSimilarity return value as:

$$\frac{1-(\mathrm{ADFA}-\mathrm{ADFSV})}{\mathrm{MADFAP}-\mathrm{ADFSV}}$$

The adjacentValueSimilarity value is calculated for each 'group' of RCDs, a group being defined as where the distance between each RCD is <1.6 Mbp, as explained in the main text. In order to provide stronger weighting for larger groups (i.e. those with larger values of n), the adjacentValueSimilarity value was logged n-1 times for final reporting. A t-test was applied to the data with and without this multiple logging.

Results

Creating a highly synchronous S phase population

In order to label DNA sequences that represent individual replicon clusters, we devised a protocol by which cells enter S phase with a high degree of synchrony. U2OS cells were synchronised using one round of thymidine treatment followed by release into normal medium and then treatment with L-mimosine[14,44–46] (Figure 1a). Release from L-mimosine allowed cells to enter S phase with a degree of synchrony significantly higher than we could obtain by release from thymidine arrest. Because L-mimosine is a hypoxia mimetic and can induce double-strand breaks[47–50], the protocol was designed to minimise the amount of time that cells were exposed to L-mimosine (Figure 1a, top panel). At different times after L-mimosine release, newly replicated DNA was labelled with 30-minute pulses of the thymidine analogue 5-ethynyl-2'-deoxyuridine (EdU); the DNA was sonicated into fragments ranging from 100–900 bp, the EdU derivatised with biotin, captured on magnetic beads and subjected to DNA sequencing (Figure 1a, bottom panel). DNA reads were then mapped back onto the reference human genome and read count was collected in 10 or 50 kb bins.

The EdU in small aliquots of cells from the different treatments was labelled with Alexa Fluor 647 for flow cytometry. Figure 1b shows 2-dimensional flow cytometry profiles showing DNA content (propidium iodide) versus EdU incorporation (Alexa Fluor 647). It took ~10 minutes after L-mimosine release for the first cells to start incorporating EdU and new cells continued to enter S phase for ~45 minutes (Underlying data: Supplementary Figure S1)[51]. Some cells remained in G1, possibly because they had not completed progression through G1 or because they had acquired double strand breaks caused by the thymidine + L-mimosine treatment. Aliquots of cells were pulsed with EdU at 10–40, 40–70, 70–100 and 100–130 minutes after L-mimosine release. Replicating cells continued to incorporate EdU at a constant rate and could be judged to be moving through S phase by an increase in total DNA content. This increase in DNA content can be seen in synchronised cells pulsed with EdU for 100-130 min (Figure 1b) where the total increase in DNA content is ~15%; shorter pulses of EdU (Underlying data: Supplementary Figure S1) also show increases in DNA content, with ~6% increase in DNA content after 40 min. This slightly slowed progression through S phase could be due to mimosine inducing a small number of double strand breaks[47–50] or reducing cellular dNTP pools[52]. Despite the synchrony protocol, some cells at different stages of S phase were still present in the whole cell population. Hence, to yield only cells labelled with EdU at the early stages of S phase we used FACS to select cells with a near-G1 DNA content (vertical red line in Figure 1b).

We examined different methods for normalising the results from DNA sequencing. To control for amplification and mappability biases, we sequenced the entire DNA content from synchronous and asynchronous cells plus or minus EdU labelling and streptavidin pull-down, as well as non-pulled-down samples. Reads from the experimental samples were normalised to the total number of counts in millions. Then, genomic background was subtracted using an approach based on CISGenome normalization[42,43] (see Methods). We tested a number of possible controls for normalisation (Underlying data: Supplementary Figure S2)[51] and decided to use for normalisation the DNA from each sample prior to pull down, hence avoiding any potential sample-to-sample variation.

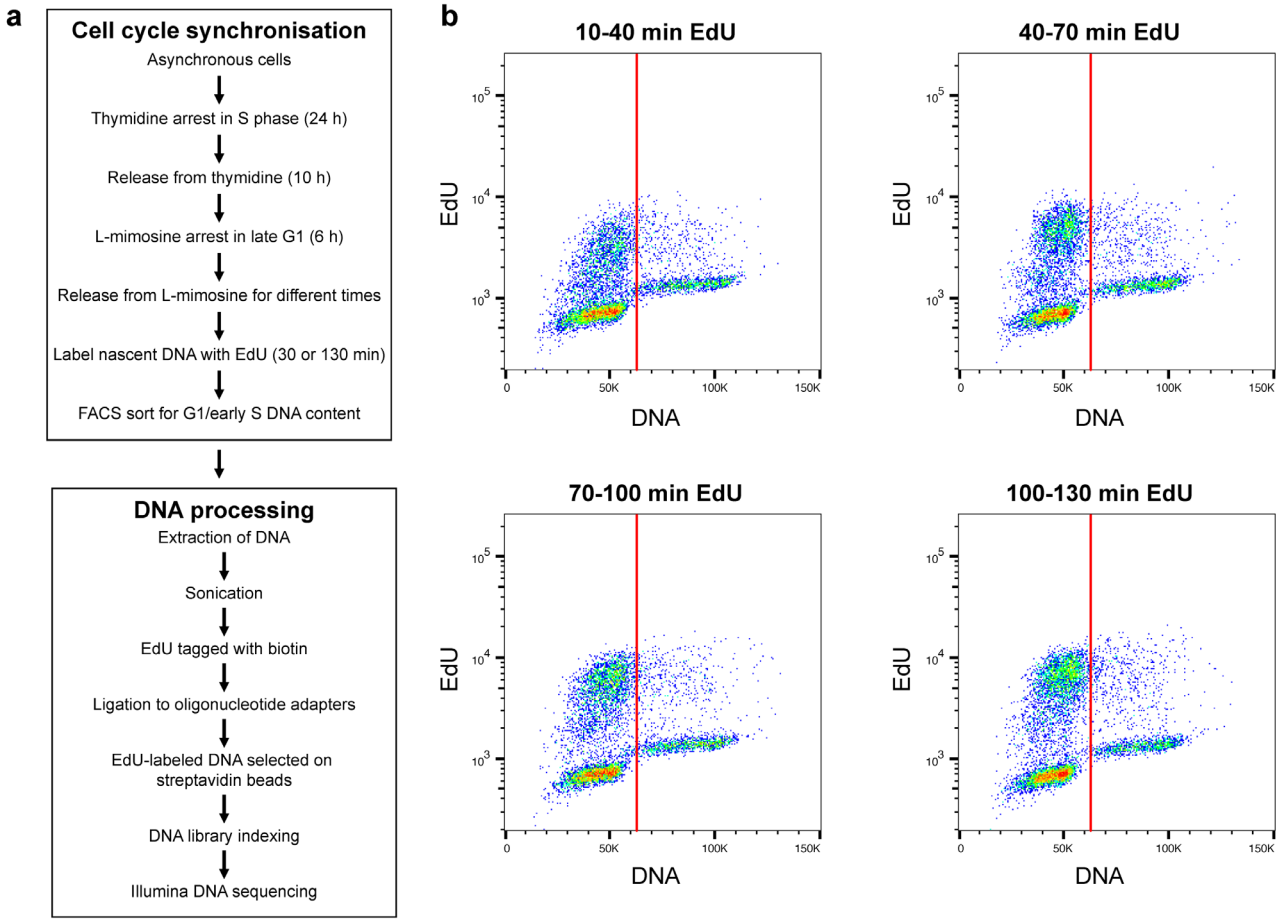


Figure 1. Synchronisation of early S phase cells. a) Upper panel: description of the procedure for synchronising U2OS cells in early S phase; lower panel: description of method for isolating replicating DNA from synchronised cells. **b**) Flow cytometry of synchronised cells. Prior to analysis cell cultures were supplemented with EdU for 30 mins at the indicated times after release from L-mimosine. EdU incorporated into DNA was labelled with Alexa Fluor 647 and total DNA was stained with Propidium Iodide. Cells were then analysed by flow cytometry: the x-axis shows DNA content (propidium iodide) and the y-axis EdU content. The red vertical lines indicate the cut-off used in preparative cell sorting experiments to include only cells with a near-G1 DNA content.

The results for chromosome 3 using this protocol are shown in Figure 2a, and data for all chromosomes is shown in Underlying data: Supplementary Figure S3[51]. Several things are apparent from this data. The replication profiles are complex, with peaks at a range of different sizes. The profiles are also remarkably consistent across the four different time points, though sharp peaks at the earlier times tend to spread out at later times. It should be noted that our sequences have been mapped onto a normal human genome, but since U2OS have a number of chromosome rearrangements and copy number variations there are a few genomic regions where there will be partial discontinuities in the DNA replication data.

We first asked how our early replication data conformed to published data on replication timing domains. Figure 2b shows the early (green) and late (red) replication timing domains for U2OS cells. There is remarkable concordance between the timing domain results and our early replication results. Genome-wide, 94% of the 10–40 min early replication signal falls into early timing domains. This is consistent with recent data on single DNA fibres which shows the high degree of stochasticity that occurs in origin firing[7]. However, the early replication peaks show much more fine structure, consistent with the idea that they represent the earliest active replicon clusters in the early timing domains.

The early replication signal can be interpreted at three different scales: the scale of individual replicons (~100 kb), replicon clusters (~500 kb) and timing domains (multi-Mbp scale). Wavelet analysis provides a means of analysing signals like this at different scales. Wavelet analysis of the 10–40 min signal is shown in Figure 2c. A wavelet of a particular width is moved across the early replication data[51] and the two signals convolved to show how much the mother wavelet is matched

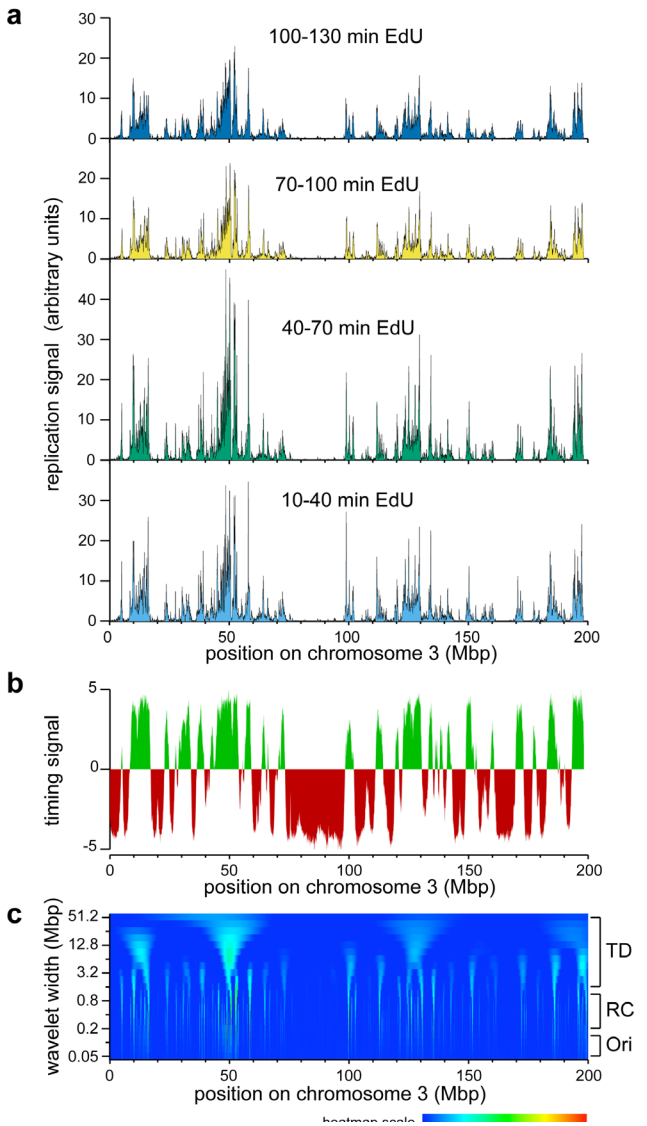


Figure 2. Early replication signals on chromosome 3. U2OS cells were synchronised in early S phase with thymidine and L-mimosine as described in Figure 1a, and were pulsed for 30 mins with EdU at different times after L-mimosine release. EdU labelled DNA was isolated and sequenced as described in Figure 1a and mapped back to the genome. EdU signals were then normalised to the respective sample's internal control reference. **a**) The normalised EdU signals on chromosome 3 are shown for the four time points. **b**) The replication timing signal for chromosome 3 from https://www2.replicationdomain.com/database.php. Early replicating DNA is shown in green and late replicating DNA in red. **c**) Heatmap (positive values only) of a wavelet analysis of chromosome 3 using a Ricker wavelet of peak width from 50 kb to 51.2 Mbp, with widths increasing by a factor of √2 between each analysis (log scale) on the y axis. At the bottom of the figure a heatmap colour scheme is shown. Dark blue represents a signal of zero or below and red represents a signal of one. The approximate size of individual replicons (Ori), replicon clusters (RC) and timing domains (TD) are indicated to the right.

at different places in the replication signal; this process is then repeated with wavelets of different widths. We chose to use a Ricker wavelet (Underlying data: Supplementary Figure S4) for the analysis because it is a simple symmetrical wavelet that is derived from a Gaussian distribution and therefore makes minimal assumptions about the expected shape of the peaks. Figure 2c shows the results when we performed this process with Ricker peak widths from 50 kb (Figure 2c bottom) to 51.2 Mbp (Figure 2c top). Features potentially corresponding to timing domains (TD), replicon clusters (RC) and individual replication origins (Ori) are visible in the heatmap.

In the rest of this paper, we first consider the potential replicon clusters in the 10–40 min time point (EdU pulse from 10–40 mins) and then consider how these replicon clusters develop over time.

Replicon cluster domains

We used wavelet analysis in order to show that the most prominent component of the early replication signal is at ~500kb in size, consistent with the expected size of replicon clusters[8,15]. To provide a detailed example of early replication, Figure 3 shows the 10–40 min signal for six 10 Mbp regions of chromosome 3 (orange bars, mapped onto 50 kb bins). Any two forks initiated from a single origin could travel 60–120 kb (2 × 30 mins × 1–2 kb/min) if they were in the first cells that entered S phase in this sample, and so could represent the finest features observed here (width of one or two orange bars). Most replicon clusters would be expected in the 300–800 kb range, and so could represent the broader peaks observable at this scale.

Figure 3 also shows heatmaps of wavelet peak width ranging from 50 kb to 9 Mbp (using data in 10 kb bins to allow analysis with the smallest wavelets). It is evident from these examples that the strongest wavelet results are with peak widths in the region of 200–800 kb, in the expected size range of replicon clusters. To determine whether this represents a global feature of early replicating DNA, we separated DNA early timing domains from later replicating DNA using the timing domain data (the green portions in Figure 2b and Underlying data: Supplementary Figure S3)[51] and then determined the mean height of all wavelet peaks in the early replication domains. Figure 4a shows this analysis for replication data in 50 kb bins and Figure 4b shows the data in 10 kb bins. We also compared DNA labelled from 10–40 mins (brown lines) with DNA labelled from 0–40 mins (blue lines). The results were similar in all analyses: the mean peak height rose rapidly from 50 kb wavelet widths to 500 kb wavelet widths, and then fell slowly. This is consistent with the visual analysis of the heatmaps (Figure 2c and Figure 3) and suggests that the most prominent features in all the early replicating regions have a width of ~500 kb, as expected of replicon clusters. On this basis we decided to use a wavelet width of

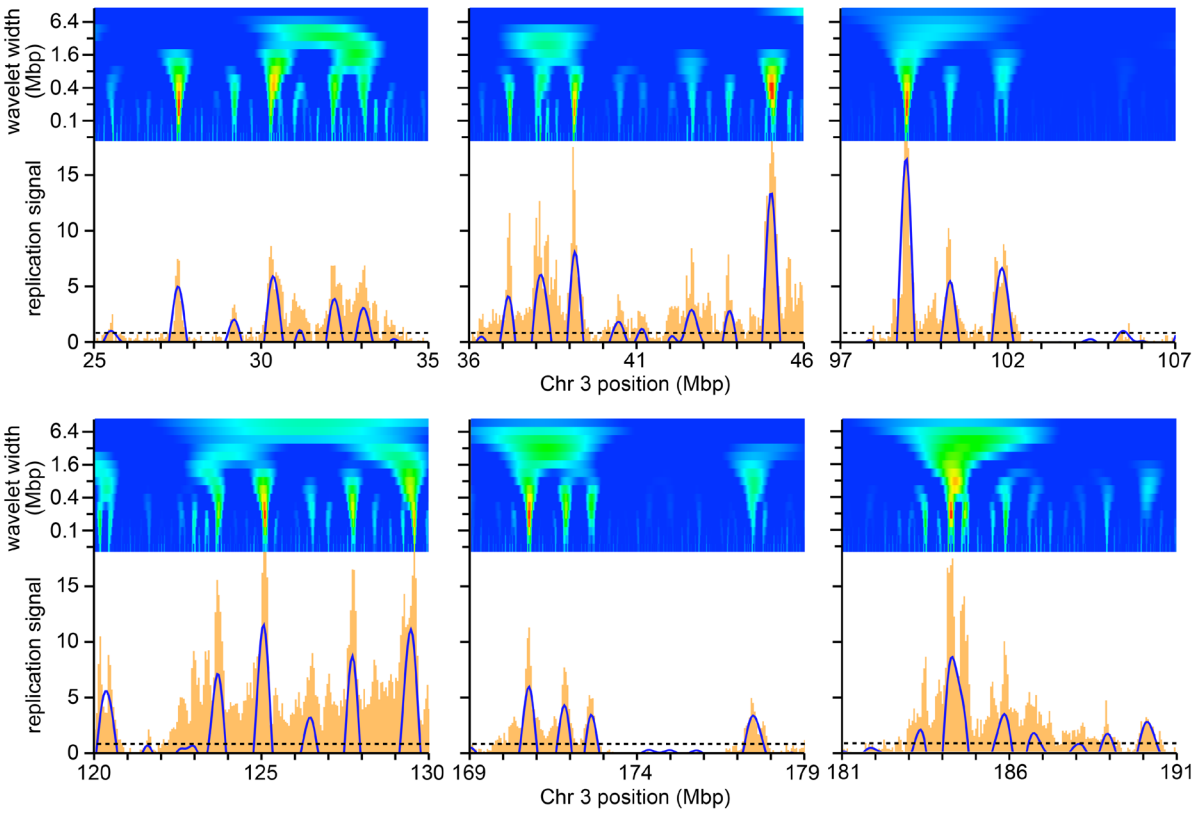


Figure 3. Illustrative results of selected regions on chromosome 3. Orange bars show the early replication signal from the first time point (EdU labelling from 10–40 mins) on six selected 10 Mbp regions of chromosome 3. Blue lines show the wavelet analysis results with a wavelet peak widths of 500 kb (positive values only). The horizontal dashed line shows the cut-off for calling a wavelet peak as set at the 70th percentile for peaks in late replicating domains (see main text for rationale). Above each replication signal is heatmap of wavelet analysis using a Ricker wavelet of peak widths ranging from 50 kb to 9 Mbp with widths increasing by a factor of √2 between each analysis (log scale) on the y axis and using data in 10 kb bins to allow analysis with the smallest wavelets. The heatmap colour scheme is the same as for Figure 2c.

500 kb to identify these features which we call 'Replicon Cluster Domains'.

The blue arcs in Figure 3 show the Replicon Cluster Domains identified with a 500 kb wavelet peak width. In order to further analyse the behaviour of these Replicon Cluster Domains, we decided to set a minimum height for calling them. We started by examining the peaks identified in the late timing domains using a wavelet width of 500 kb (the red portions in Figure 2b and Underlying data: Supplementary Figure S3)[51] as this largely consists of features that look like background levels of EdU incorporation. We explored the effect of setting the cut-off for peak recognition at different percentiles of the peak heights in the late timing domains. Figure 4c shows the effect on Replicon Cluster Domain identification in early timing domains of different percentile cut-offs. There were a total of 1262 wavelet peaks in early timing domains (wavelet width 500 kb). Setting a minimum cut-off at the 70th percentile of peaks in late replicating domains (i.e., removing 70% of the peaks in late domains) removed 122 (9.7%) of the smallest peaks in the early timing domains, leaving 1140 Replicon Cluster Domains. Cut-off percentiles >70% started to significantly reduce the number of Replicon Cluster Domains in the early domains. Figure 4d shows the distribution of wavelet peak heights in the early timing domains and the effect of implementing the 70th percentile cut-off. Establishing this 70th percentile cut-off did not significantly change the 500 kb optimum wavelet peak width (dashed lines in Figures 4a and 4b). The 70th percentile cut-off value is shown as a dashed horizontal line in Figure 3.

Figure 5 shows some metrics of the 1140 Replicon Cluster Domains identified in this manner. Figure 5a shows that the peak separation between Replicon Cluster Domains is fairly uniform, with a mean of 1.187 ± 0.43 Mbp. Since the width of each Replicon Cluster Domain is about 0.5 Mbp, this means that Replicon Cluster Domains peaks are quite closely packed together in early timing domains. Figure 5b shows that there is a clear proportionality between the size of a timing domain (x axis) and number of Replicon Cluster Domains it contains (y axis). This means that there is a fairly close packing of Replicon Cluster Domains within all early timing domains.

The >20-fold variation in the height of the replication peaks representing the Replicon Cluster Domains (Figure 4d) is striking. Because our sequencing results were carefully normalised to total genomic DNA (Underlying data: Supplementary Figure S2)[51] higher peaks indicate more DNA replication (EdU) in that region. The difference in peak heights could be caused by a combination of two different effects: it could represent cell-to-cell variation in the densities of active origins in each Replicon Cluster Domain or it could represent some stochasticity in the time that Replicon Cluster Domains become active. From the extensive data on replicon sizes obtained by DNA fibre analysis it seems highly unlikely that variability in the density of replication origins could account for all of the peak height variability. Since DNA fibre analysis also suggests that the vast majority of origins fire near-synchronously in clusters, we conclude that the variability in peak height implies that there

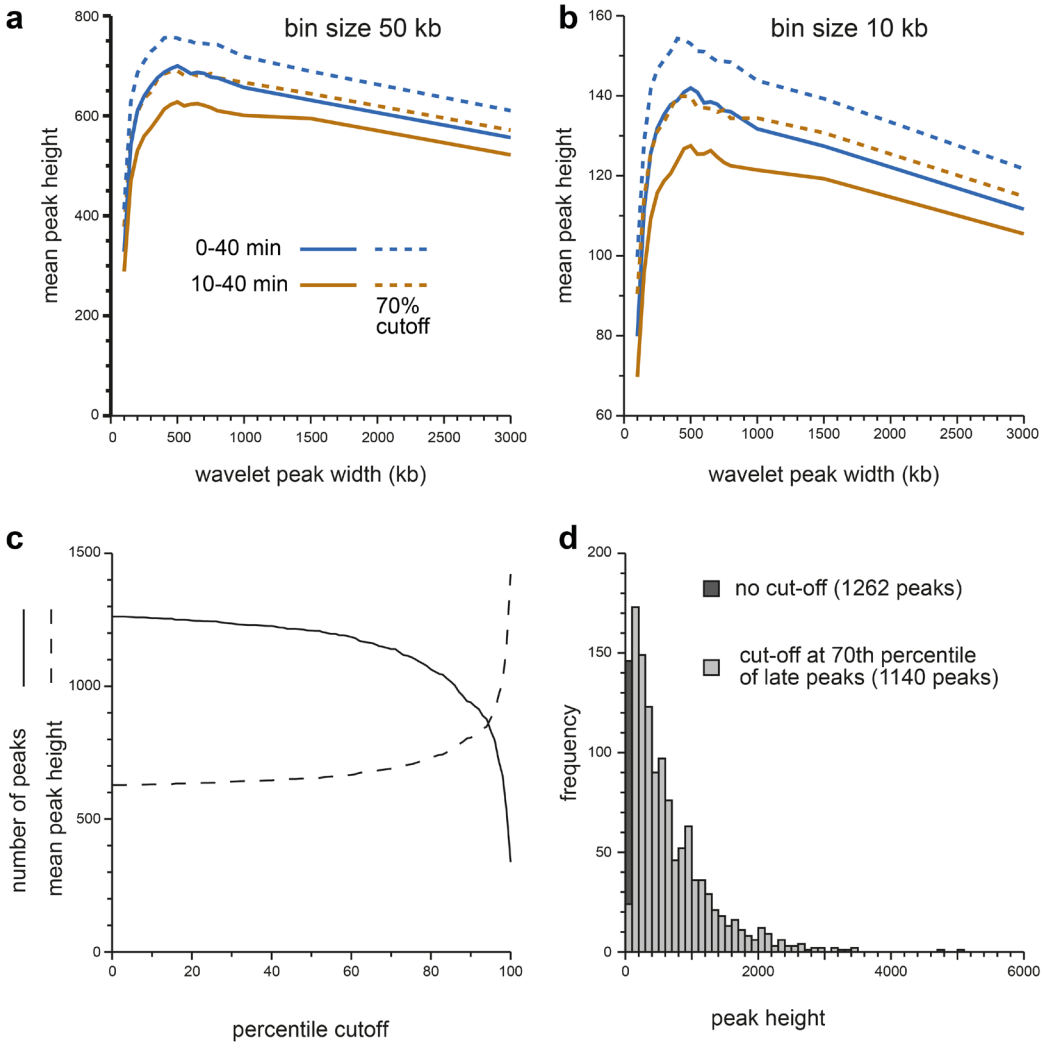


Figure 4. Metrics from a genome-wide wavelet analysis. a, **b**) The early replication signals from the first time point (EdU labelling from 10–40 mins, brown lines) or an extended first time point (EdU labelling from 0–40 mins, blue lines) with reads mapped onto 50 kb bins (panel **a**) or 10 kb bins (panel **b**) were clipped to include only the early timing domains and were analysed using a range of closely-spaced wavelets (peak widths of 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 1000, 1500 and 3000 kb). Peaks were then called optionally using a minimum peak hight representing the 70th percentile for peaks in late replicating domains (solid lines without cutoff, dashed lines with cutoff). The height of each peak was recorded and the total genome-wide sum for each data point is plotted. **c**) The early replication signal from the first time point (EdU labelling from 10–40 mins) was clipped to include only the late timing domains and analysed using a wavelet with 500 kb peak width. The peak heights were sorted in order and expressed as a percentile. The early replication signal from the first time point (EdU labelling from 10–40 mins) was then clipped to include only the early timing domains and analysed using a wavelet with 500 kb peak width. Peak calling was then performed using as a cut-off the wavelet peak height at different heights derived from the late-replicating DNA. Solid line shows the number of wavelet peaks (solid line) and the mean wavelet peak height (dashed line) called in the early timing domains using different percentile cut-offs derived from the late timing domains. **d**) The distribution of wavelet peak heights obtained from applying a wavelet of 500 kb to the first time point (EdU labelling from 10–40 mins) clipped to include only the early timing domains. The darker bar shows the effect of using a 70th percentile cutoff derived from the replication signal in late timing domains, which removes 87 of the smallest wavelet peaks.

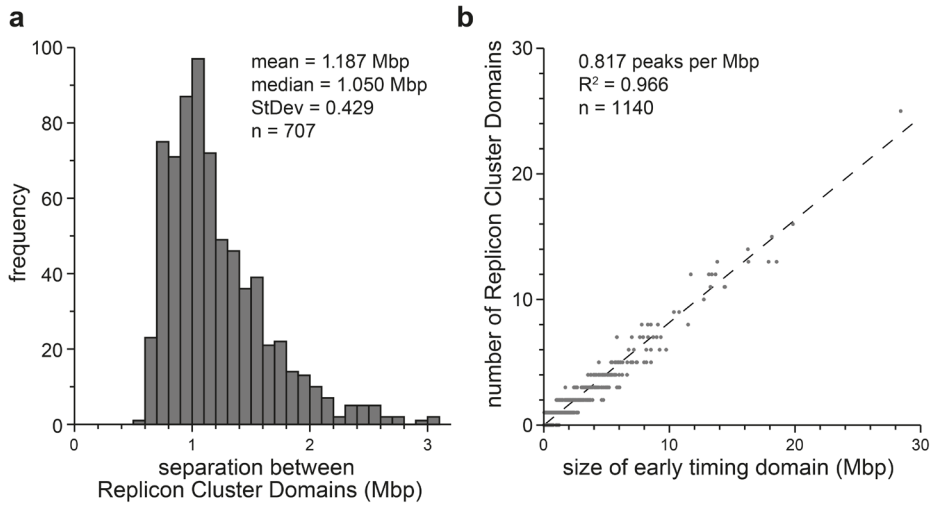


Figure 5. Separation between Replicon Cluster Domains. A wavelet of width 500 kb was applied to the first time point (EdU labelling from 10–40 mins) clipped to include only the early timing domains (data mapped in 50 kb bins). Wavelet peaks falling below the 70th percentile of peaks identified in late timing domains were removed. The remainder of the peaks were classified as Replicon Cluster Domains. **a**) The distance between adjacent Replicon Cluster Domains identified within each early timing domain was recorded. The frequency distribution of the separation between adjacent peaks is shown. **b**) The number of Replicon Cluster Domains identified in each early timing domain is plotted against the size of the early timing domain.

is considerable cell-to-cell variation in the time that Replicon Cluster Domains become active in different cells. We will explore this idea later.

Growth of individual RCDs during S phase

We next turned our attention to understanding how individual Replicon Cluster Domains develop over time. The movement of the two forks at the outer edges of each replicon cluster will expand the margins of each Domain at successive time points. Visual inspection of the replication data (Figure 2a and Underlying data: Supplementary Figure S3)[51] suggests that in passing through the four different time points, most peaks stay in the same position but get wider. In order to investigate this systematically and quantitatively, we used a simple algorithm to identify individual peaks that are sufficiently separated from their neighbours that a measurement of their width can be made. To do this, we started with the 1140 Replicon Cluster Domains identified as in Figure 5, using the 10–40 minute EdU data analysed by a wavelet peak width of 500 kb and implementing a 70th percentile cut-off relative to peaks in late timing domains. We then removed any Replicon Cluster Domains where the total amount of replication signal in the 500 kb on either side of the wavelet peak was more than 25% of the replication signal under the wavelet peak. This gave 123 'isolated' Replicon Cluster Domains in the 10–40 minute EdU data (listed in Underlying data: Supplementary Figure S5)[51]. Figure 6 shows examples of six of these isolated Replicon Cluster Domains as they develop over the four time points. A broadening and flattening of the peaks over time is evident in these examples.

Figure 7a and 7b shows how idealised replicon clusters might look if they were pulsed with EdU over five successive time points. Because different origins may be selected to fire in different cells, the population will have curved (Gaussian-looking) edges, but because origin density within clusters is somewhat uniform, the curve will have a flatter top than a Gaussian curve. Origins within a single cluster fire near-synchronously, so that the flat-topped nature of the peak will increase with time. Once all internal forks have terminated, EdU incorporation will occur only at the outer edges of the cluster. The peak shapes in Figure 6 conform to some degree with the model in Figure 7a, though there is little evidence for the termination-driven peak-splitting, which suggests that most Replicon Cluster Domains are still replicating internally even at the last time point.

We analysed all 123 isolated domains using a range of wavelets with peak widths at 25 kb intervals from 200 kb to 1,200 kb and determined the best fit to the experimental data. We rejected from further analysis 51 peaks that fitted to the extreme values of 200 kb or 1,200 kb at any of the four timepoints. The location and optimal wavelet width of the remaining 72 isolated peaks is given in Underlying data: Supplementary Figure S5[51] and they are shown in blue for the exemplar peaks in Figure 6.

As an alternative way of measuring the widths of isolated peaks, we took the same 123 isolated domains and fitted Gaussian curves to them using a Nelder-Mead algorithm. We rejected from further analysis 26 peaks that fitted to the extreme values of 200 kb or 1,200 kb at any of the four timepoints. The location and full width of maximum height of the remaining 97 isolated peaks is given in Underlying data: Supplementary Figure S5[51] and they are shown in red for the exemplar peaks in Figure 6.

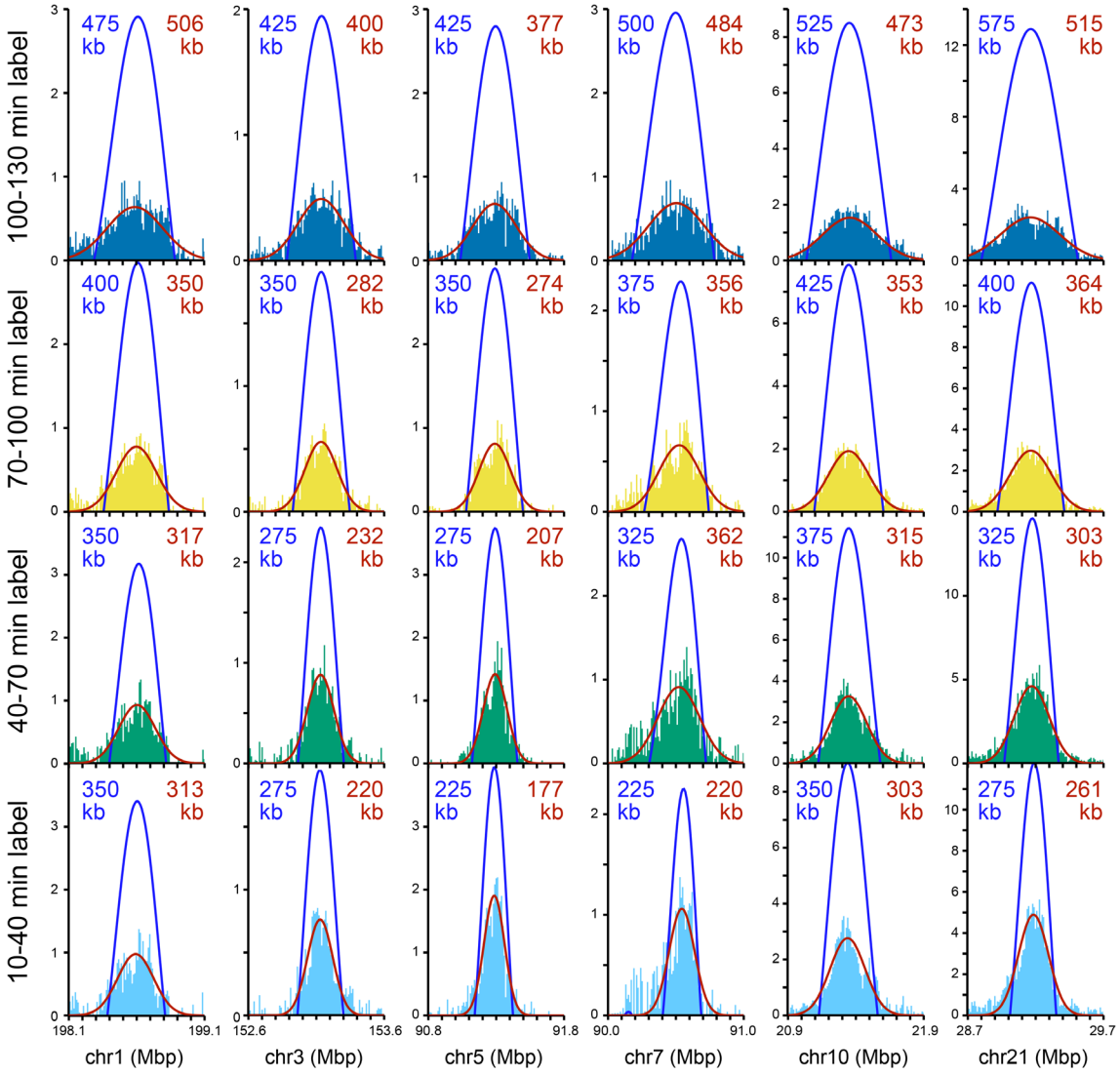
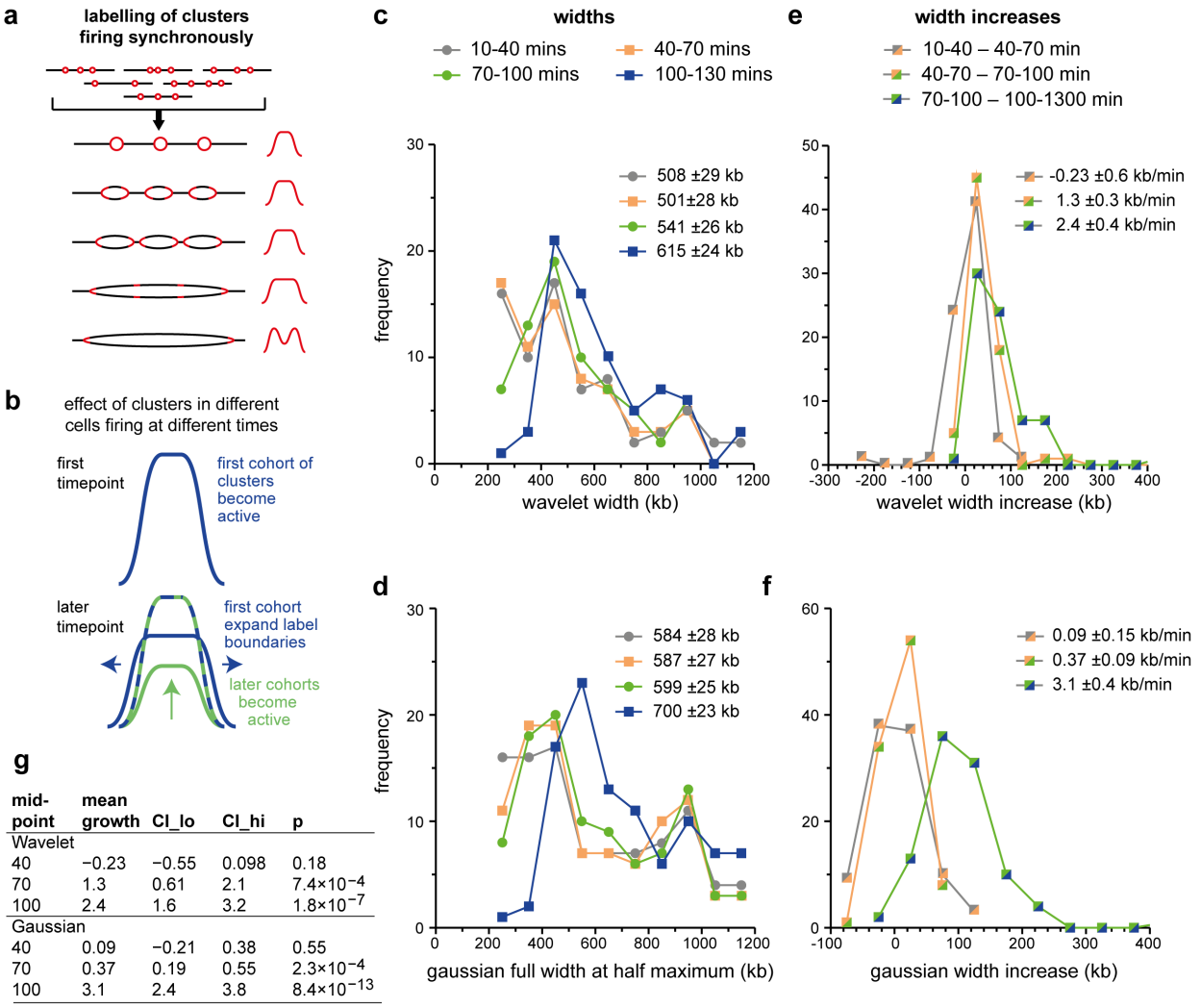


Figure 6. Exemplar isolated Replicon Cluster Domains. Six Replicon Cluster Domains that were substantially isolated from other replication signals were chosen as exemplars. Their replication signal in the four time points is shown (light blue, green, yellow and dark blue bars). Each Replicon Cluster Domain was analysed by a series of finely separated wavelets (in 25 kb intervals from 200 kb to 1,200 kb; data in 10 kb bins). The optimal wavelet selected is shown in blue and its width shown in blue text. The same six peaks were fitted to a Gaussian curve which is plotted in red; the red text shows the full width at half maximum of the curve.

Figure 7c shows the width of the fitted wavelets, which increase in size from an average of 508 kb in the 10–40 min time point to 615 kb in the 100–130 kb time point. Figure 7d shows similar data for the width of the fitted Gaussian peaks, which increase from an average of 584 kb to an average of 700 kb. Figures 7e and 7f show how the width of each isolated peak increases in size between successive time points as analysed by wavelet fitting (7e) or Gaussian curve fitting (7f). There is no significant increase in average wavelet width between the first two time points, but between the second and third time point widths increase by an average of 40 kb (wavelet) or 12 kb (Gaussian) and between the third and fourth time points widths increase by an average of 73 kb (wavelet; 2.4 kb/min) or 101 kb (Gaussian; 3.1 kb/min). Figure 7g shows that isolated peaks grow in width at later stages (> 70 min) at increasing rates of 0.4 - 3 kb / min in a statistically significant manner.

Replication forks in early S phase U2OS cells move at 1–1.5 kb/min[9], so the two flanking forks in a replicon cluster would be expected to expand the width of the peak by 60–90 kb in successive labelling periods. This is in line with the later expansion rates. The lack of significant growth between the first two time points (between the 10–40 mins and the 40–70 mins timepoints) can be explained by cells continuing to enter S phase over ~45 minutes after L-mimosine



g

mid-point	mean growth	CI_lo	CI_hi	p
Wavelet				
40	−0.23	−0.55	0.098	0.18
70	1.3	0.61	2.1	7.4×10^{-4}
100	2.4	1.6	3.2	1.8×10^{-7}
Gaussian				
40	0.09	−0.21	0.38	0.55
70	0.37	0.19	0.55	2.3×10^{-4}
100	3.1	2.4	3.8	8.4×10^{-13}

Figure 7. Development of isolated Replicon Cluster Domains. a) Schematic of an idealised Replicon Cluster Domain containing three active origins. Because of stochasticity in the selection of active origins plus near synchronous initiation within the cluster, as shown in the cartoon to the left, the EdU incorporation profiles on the right would tend to have a flat top and curve down at the edges. The width of the EdU incorporation profiles will expand driven by the outer forks until all internal forks have terminated and incorporation is restricted to the two outer forks. **b**) Schematic of the EdU incorporation profiles in a single Replicon Cluster Domain in the population of synchronised cells. The green line shows cells where the RCD had become active only in the current timepoint, so its margins more closely match the edges of the RCD. The blue line shows cells where the RCD had become active in a previous timepoint and so its edges will expand due to forks expanding at the margins. The blue-green line shows the observed EdU signal, which comprises elements of both the green and the blue labelling. **c**-**f**) The Replicon Cluster Domains from the 10–40 minute EdU data as defined in Figure 5 were filtered to exclude those where the total amount of replication signal in the 500 kb on either side of the wavelet peak was >25% of the replication signal under the wavelet peak. These 123 'isolated' peaks were fitted to an optimal wavelet (using a range of wavelets with peak widths at 25 kb intervals from 200 kb to 1,200 kb; panels **c** and **e**) or Gaussian curve (panels **d** and **f**) over the four time points. Peaks that fitted to the extreme values of 200 kb or 1,200 kb at any of the four timepoints were rejected. The optimal wavelet width (panel **c**) or the Gaussian width (panel **e**) are plotted. The width change between successive time points is plotted in panel **d** (wavelet) and panel **f** (Gaussian). The mean width and their standard error for each time point is also listed in panels c and d. The mean width increase between successive times point and its standard error is also listed in panels **e** and **f**. **g**) Statistical analysis of width increases. Mean growth is given in kb/min. CI lo and CI hi are a 95% confidence interval of the mean. The p value is the result of a one sample t-test against zero for peak width increases.

release (Figure 1) which will tend to minimise the mean width of the labelled peak, as portrayed in Figure 7b.

A typical replicon within a replicon cluster might remain active for 37–75 min[4,8–11]. Visual inspection of the isolated peaks showed only a few examples of the peak splitting that would be expected to occur once all internal forks have terminated (as depicted in the last cartoon in Figure 7a). This suggests that even in the last time point (100–130 mins EdU) most Replicon Cluster Domains are still being replicated by internal forks in some cells in the population.

Replication of valleys

We next considered the development of Replicon Cluster Domains packed together within large timing domains. Figure 8 shows three representative 10 Mbp segments that contain multiple Replicon Cluster Domains. The bottom of the figures shows published data on replication timing domains (early in green, late in red). The overall shape of the peaks was maintained over the two-hour time course, but there was a clear widening of the peaks as the valleys between the peaks became filled in. We considered whether this represents simple expansion of peaks due to the movement of its two flanking forks (as depicted in Figure 7a) or whether it is driven by new initiation events. Brown bars were drawn over peaks identified by wavelet analysis of the first time point and at successive time points these bars were lengthened by 90 kb (fork rate 1.5 kb/min).

The data show that in some regions where Replicon Cluster Domains are closely spaced the brown bars have merged by the last time point, consistent with the idea that some valley-filling could be accounted for by fork progression at the edges of clusters. However, many gaps between the brown bars still remain in the last time point so it is clear that this cannot account for replication of all the valley DNA. The median separation between Replicon Cluster Domains is ~1.2 Mbp (Figure 5) and if they have a width of ~500 kb in size flanking forks progressing ~360 kb over the two-hour time course only extend the median cluster to a width of ~860 kb, which is two thirds of the distance required. In addition, we showed in Figure 7 that fork-driven expansion of the edges of Replicon Cluster Domains is not clearly seen in the first time point due to the continued entry of cells into S phase. This suggests that complete replication of valley DNA within a two hour period would will depend on further initiation events.

There is also good evidence for initiation in the valleys between Replicon Cluster Domains from the replication data (Figure 2, Figure 3, Figure 8 and Underlying data: Supplementary Figure S3)[51] which shows that most valleys begin to replicate, albeit to a low level, even in the first time point. Because of the relatively high degree of synchrony obtained in these experiments (Figure 1) we do not believe this 'valley labelling' is due to contaminating cells at a slightly later stage of S phase.

Instead, we favour the idea that there are active replication origins within these valleys. This would also be consistent with the considerable variation in the height of the peaks representing the Replicon Cluster Domains (Figure 2 and Underlying data: Supplementary Figure S3)[51] which could be caused by them becoming active at different times, with lower peak heights representing Domains that tend to become active at later times. Since DNA fibre analysis suggest that the majority of origins fire in clusters, our data would be consistent with the idea that 'valleys' between the prominent Replicon Cluster Domains also represent Replicon Cluster Domains that tend to become active at later times and which may also be invaded by forks emanating from neighbouring

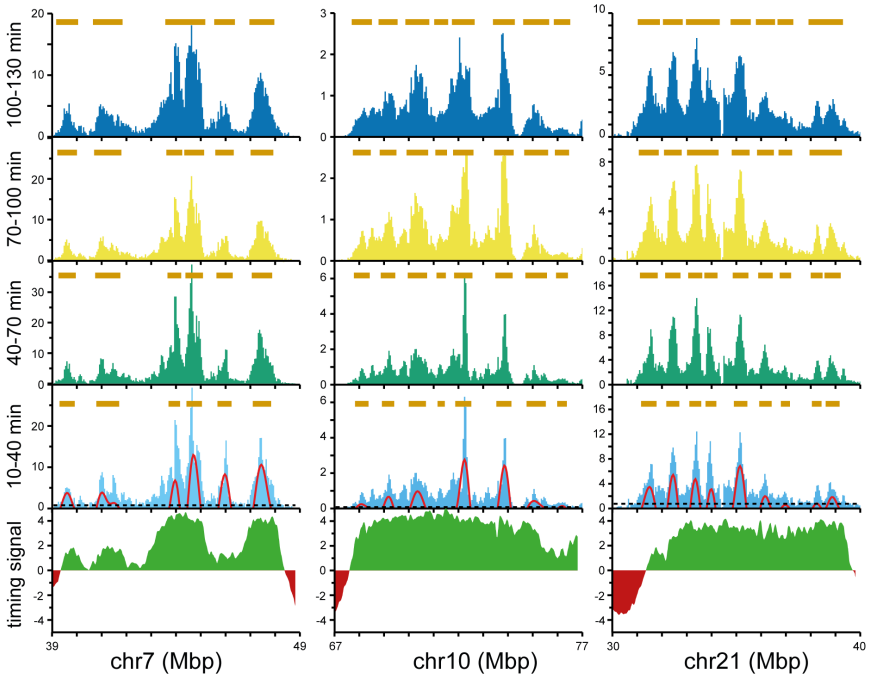


Figure 8. Examples of valley-filling between Replicon Cluster Domains. The early replication signal of three selected regions at the four time points is shown as light blue, green, yellow and dark blue bars (data in 50 kb bins). For the first time point (10–40 mins EdU) the wavelet analysis results with a wavelet peak width of 500 kb are shown by the red lines. The horizontal dashed line shows the cut-off for calling a wavelet peak as set at the 70th percentile for peaks in late replicating domains. For the first time point, wavelet peaks are marked by horizontal brown bars; for the successive timepoints the edges of the bars were extended by 90 kb as expected of a fork moving at 1.5 kb/min. At the bottom the relevant timing domain signals for the regions are shown (early timing domains in green and late timing domains in red).

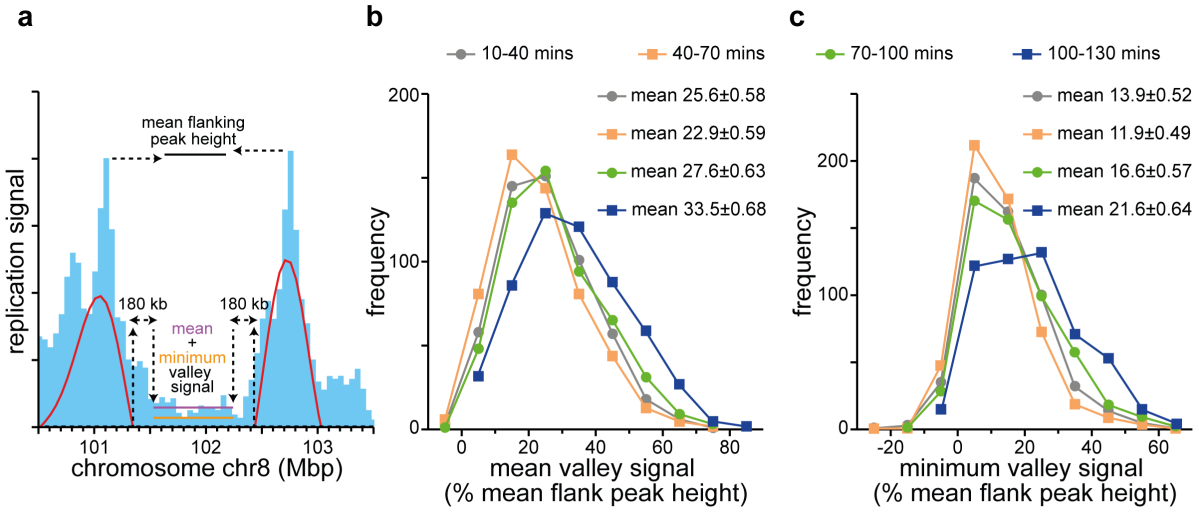
Replicon Cluster Domains that have become activated earlier in S phase.

To explore this idea further, we analysed the evolution of 'valleys' in a more systematic way. Figure 9a shows as an example a single valley on chromosome 8. We started with the 500 kb wavelet analysis of the first time point (10–40 mins EdU) though without any height cut-off for calling the peaks. As shown in Figure 9a, we defined 'valleys' as regions between two pairs of adjacent wavelet peaks residing in a single timing domain after we had added 180 kb to the edges of the wavelet peaks to account for fork movement (1.5 kb/min over 120 minutes). Six hundred and forty 'valleys' conformed to this definition in the first time point. We then examined the replication signal in the three later time points and rejected any valleys where the flanking peaks had shifted by more than 100 kb; this left 540 valleys that could be tracked across all four time points. Because the replication signal is not normalised between the different time points, we expressed the mean and minimum signal in each valley as a percentage of the mean height of the flanking peaks (Figure 9a).

The frequency distribution of mean and minimum valley signals for the four timepoints are shown in Figures 9b and 9c respectively. Figures 9d shows that there is a statistically significant increase in both mean and minimum valley filling at later time points. Importantly, by the last time point virtually every location in every valley has incorporated a significant amount EdU (525 out of 540 valleys have their minimum signal >10% mean flanking peak signal). This valley replication most likely represents new initiation events and is consistent with the idea that RCDs become activated right across the early timing domains, though with valley RCDs tend to activate later.

Sequential activation of RCDs

There is experimental support for the idea that the different replicon clusters comprising a single timing domain are activated sequentially, the so-called domino model[30,53–56]. In some DNA fibre studies that have analysed very long stretches of DNA, a second cluster has been shown to become active after a first cluster has been activated[55]. In addition, dual labelling of replication foci shows evidence for a 'domino' activation model, where a second replication focus becomes activated



d

	mean midpoint	filling rate (% / min)	CI low	CI high	p value
mean	40	−0.025	−0.034	−0.016	7.2×10^{-8}
	70	0.087	0.077	0.096	6.7×10^{-56}
	100	0.20	0.19	0.21	6.8×10^{-149}
min	40	0.036	0.024	0.048	5.9×10^{-9}
	70	0.052	0.04	0.065	9.7×10^{-15}
	100	0.16	0.15	0.18	1.1×10^{-79}

Figure 9. Analysis of valley-filling genome-wide. a) Schematic of how 'valleys' were analysed genome-wide, using a region on chromosome 8 as an example. The replication signal in the first time point (10–40 mins EdU; 50 kb bins) was subject to wavelet analysis with a wavelet peak width of 500 kb (red lines). For all pairs of wavelet peaks within any given early timing domain, the internal edges of the gap between the edges of the wavelet peaks were reduced by 180 kb to account for fork movement at 1.5 kb/min over two hours; the remaining gap was defined as the 'valley'. Wavelet analysis with a peak width of 500 kb was then performed on the three later time points, and only those valleys whose flanking wavelet peaks existed (±100 kb error) in all four time points were included for further analysis. For each time point, the mean valley replication signal (horizontal purple line) and minimum valley replication signal (horizontal orange line) were expressed as a percentage of the mean height of the replication signals of the two flanking peaks (black horizontal line). **b**) The frequency distribution of the mean valley replication signal across the four time points. The mean and standard error of the distribution is also given. **c**) The frequency distribution of the minimum valley replication signal across the four time points. The mean and standard error of the distribution is also given. **d**) Statistical analysis of filling rates. Filling rates are given in %/min. CI low and CI high are a 95% confidence interval of the mean. The p value is the result of a one sample t-test against zero.

immediately adjacent to a previously activated focus[53,54]. The mechanism by which this occurs is not known, however, it is unlikely to occur by the forks emanating from one domain stimulating initiation of the adjacent domain because: a) disabling the S phase checkpoint in the absence of processive elongation of replication forks still results in the sequential firing of replication foci in the proper order[57] and; b) optical replication mapping studies failed to find evidence for new bursts of DNA synthesis near longer tracks[7].

We therefore investigated whether we could see evidence for an ordered domino-type activation of Replicon Cluster Domains in our early replication data; by this we mean that whether adjacent RCDs replicate sequentially at a frequency higher than random. A cursory look at the time course data (Figure 2, Figure 3, Figure 8 and Underlying data: Supplementary Figure S3)[51] shows that Replicon Cluster Domains packed together in groups are not strictly arranged in order of height, indicating that there is no absolute 'domino order' by which they are activated. Nevertheless, we decided to investigate in a systematic way whether there is any evidence for a preferential activation order of adjacent Replicon Cluster Domains.

We first developed a metric to select 'groups' of adjacent Replicon Cluster Domains that might display some sort of activation order. Within individual timing domains, the mean distance between adjacent Replicon Cluster Domains is 817 kb (Figure 5b). We therefore defined adjacent Replicon Cluster Domains as being in the same 'group' if the distance between them was less than twice this value, i.e., <1.6 Mbp. The height of the peak was then defined as the maximum replication signal within the Replicon Cluster Domain, using the replication data mapped onto 50 kb bins. The activation order, as inferred from peak height, was analysed in two different ways. For small groups of three or four Replicon Cluster Domains, we examined every different permutation of peak height ranking. We next used an 'adjacent height similarity' metric to provide a global view of all groups of Replicon Cluster Domains.

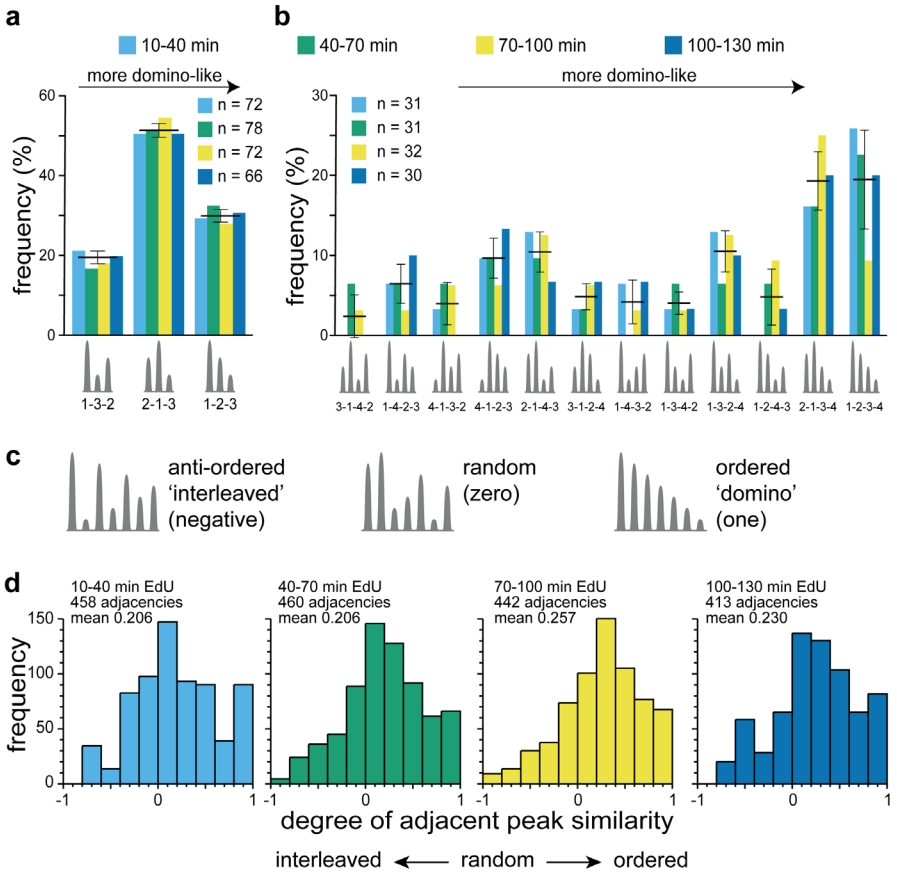
For groups containing three or four peaks we classified the height order of the peaks within each group, with the highest given a value of one, the next highest two and so on and then considered all the possible permutations of height order (bearing in mind that direction along the chromosome is arbitrary). The results are shown in Figure 10a (groups of three Replicon Cluster Domains) and 10b (groups of four Replicon Cluster Domains). All permutations are represented in the experimental data. However, some permutations are more abundant than others. For groups of three Replicon Cluster Domains there are three possible permutations with each having an expected frequency of 33.3% if the order was random. The data show a marked bias in the permutations, with the 2-1-3 ordering being the most abundant at 51.4 ± 1.7% across the four timepoints and the 1-3-2 ordering being the least abundant at 18.8 ± 1.6%. This bias towards the 2-1-3 ordering is consistent with there being a preference to activate a new Replicon Cluster Domain adjacent to a previously active one. For groups of four Replicon Cluster Domains there are twelve possible permutations with each having an expected frequency of 8.3% if the order was random. Again, the two most represented permutations are ones with the highest degree of height similarity: 1-2-3-4 (at 19.4 ± 16.2%) and 2-1-3-4 (at 19.3 ± 13.6%). These results are consistent with a 'domino' activation sequence being preferred though not strictly necessary.

To test this against all group sizes, we devised an 'adjacent height similarity' metric which reported whether the peaks were in perfect height order ('domino') with a value of 1, randomly ordered with a value of 0, or avoided adjacent height similarity ('interleaved') with a negative value (Figure 10c). This metric was calculated for all groups of Replicon Cluster Domains containing four or more members and the distributions are shown in Figure 10d. At all four timepoints there was a clear quantitative preference for peaks of similar sizes to lie next to one another. Figure 10c shows that there is a strong statistical preference for peaks being activated in sequence across a group. It is not possible to tell from the population data whether this preference is seen within individual cells or whether it is only a feature of the population as a whole. However, the result is consistent with the idea that whilst the activation timing of Replicon Cluster Domains is variable in the population, within individual cells the presence of an active Replicon Cluster Domain enhances the probability that an adjacent Replicon Cluster Domain subsequently becomes active.

Discussion

We provide here the first evidence that we are aware of for the existence of Replicon Cluster Domains at a DNA sequence level. We have labelled and sequenced replicating DNA from cells passing through the early stages of S phase at a relatively high degree of synchrony compared to previous studies. The replication signal in our data is consistent with replication timing data from cells at a lower degree of synchrony, but our data show a much higher degree of fine-scale structure. Our results provide a bridge between results obtained by DNA fibre analysis and microscopy with mapped genomic loci.

Although the existence of replicon clusters – adjacent groups of synchronously firing origins – has been known for a very long time, the genomic sequences that they correspond to have remained unknown. We report here that early replicating DNA shows a broad range of peaks at specific and consistent genomic locations. The genomic locations of these early replicating peaks are highly consistent with previous results using a lower temporal resolution and fall within regions previously identified as Early Timing Domains[22,26,27]. However, our results show much more fine-scale structure than was shown by previous studies analysing the structure of Replication Timing Domains. This is consistent with the idea that individual Replication Timing Domains consist of a collection of smaller domains that become active at slightly different times, either early or late in S phase[22,30]. It has previously been suggested that Replication Timing Domains consist of smaller functional units because the locations of Replication Timing Domains vary between different cell types, and differences in Timing Domain location tend to divide them at reproducible positions.



e

time point	mean	CI low	CI high	p value	no-rep p value
10-40	0.21	0.17	0.25	6.1×10^{-22}	6.1×10^{-22}
40-70	0.21	0.17	0.25	1.5×10^{-22}	2.0×10^{-22}
70-100	0.26	0.22	0.30	3.1×10^{-31}	1.2×10^{-31}
100-130	0.24	0.19	0.28	2.4×10^{-25}	4.8×10^{-25}

Figure 10. Activation order of Replicon Cluster Domains. The replication signal in the four time points (50 kb bins) was subject to wavelet analysis with a wavelet peak width of 500 kb. Wavelet peaks falling below the 70th percentile of peaks identified in late replicating DNA were removed. The peak height of the replication signal within each wavelet peak was recorded. Groups of peaks were defined by being in the same timing domain and being less than 1600 kb apart. **a**) Analysis of groups with three members. Each peak was given a rank order dependent on the height of its replication signal and was classified into one of three possible height order permutations (1-2-3, 1-3-2 or 2-1-3). The percentage of groups with each activation sequence over the four time points is shown. Bars showing mean and standard deviation across the four timepoints are also shown. The permutations have been ordered so that moving left to right are permutations that are more 'domino-like' and have increasing height similarity. **b**) Analysis of groups with four members. Each peak was given a rank order dependent on the height of its replication signal and was classified into one of twelve possible height order permutations. The percentage of groups with each activation sequence over the four time points is shown. Bars showing mean and standard deviation across the four timepoints are also shown. The permutations have been ordered so that moving left to right are permutations that are more 'domino-like' and have increasing height similarity. **c**) Schematic showing possible examples of height order groups with 7 peaks. If adjacent peaks have maximal height differences, they will display an interleaved pattern and have a negative score in the adjacent peak similarity metric. If peak heights are randomly distributed they will on average have a zero score in the adjacent peak similarity metric. If peak heights are in perfect ('domino') order they will have a score of one in the adjacent peak similarity metric. **d**) Groups of peaks with four or more members were analysed by the adjacent peak similarity metric, and the frequency distribution in the four different time points is shown. Groups with three members were omitted from this analysis because with only three members the metric cannot distinguish between random and anti-ordered. **e**) Statistical analysis of the adjacent peak similarity metric. CI low and CI high are a 95% confidence interval of the mean. The p value is the result of a one sample t-test against zero. The 'no-rep' p value is a similar one sample t-test against zero where each group of peaks was logged only once (see Methods for details).

Wavelet analysis of the early replicating DNA shows that the maximum signal for these peaks is 500 kb, almost exactly as would be expected from quantitative analysis of DNA fibre analysis of replicon clusters[8,15,55]. As cells progress through S phase, the width of these peaks grows at rates consistent with the progression of a single pair of replication forks at the extreme edges of each peak. We therefore propose that these peaks represent the DNA synthesised by individual replicon

clusters and have named their genomic locations 'Replicon Cluster Domains'. Since there is widespread speculation that replicon clusters might correspond to replication foci[8,15,30,36], Replicon Cluster Domains might also correspond to the genomic positions where replication foci are active early in S phase. Boundaries between Replicon Cluster Domains might also represent the boundaries between Topologically Associated Domains which have been shown to represent the borders between replication timing domains conserved between different cell lines[23,58].

One striking feature of the peaks of early replicating DNA is an extreme variability in their heights spanning a >20-fold difference. Whilst some degree of variation in the intensity of the replication signal would arise from differences in the density of active replication origins, this degree of difference in origin density has not been observed by classical DNA fibre analysis where individual DNA fibres are selected and individually analysed[8,15]. However, our results appear to be consistent with the extreme stochasticity recently observed using a high-throughput single-molecule approach[7]. The alternative explanation for the height differences is that that there is some stochasticity in the time that Replicon Cluster Domains become active, and the height of a peak reflects the number of cells in the population that have an actively replicating cluster at that particular location. By this interpretation, which we favour, high peaks represent Replicon Cluster Domains that are efficiently activated early in S phase, and low peaks represent Replicon Cluster Domains that are inefficiently activated early in S phase and/or typically become activated only later in S phase.

Replicon Cluster Domains are packed fairly tightly into early replicating DNA with mean spacing of 1.2 Mb across all early replication timing domains. However, there is no regularity in this spacing, consistent with the idea that Replicon Cluster Domains vary in size and can become active at different times in different cells in the population. This conclusion is also supported by examination of the replication of the valleys between the early replicating peaks. In the two-hour time course we examined, only a few of the valleys could be completely replicated by the forks from the edges of the flanking peaks, suggesting the need for further initiation events to complete replication. Consistent with this idea, we see the replication signal building up with time at locations too far from the flanking peaks to be accounted for by replication forks progressing outwards from the peaks. The mean spacing of 1.2 Mb early-firing Replicon Cluster Domains is consistent with the presence of Replicon Cluster Domains in the valley bottoms that are activated later in S phase. Taken together, our data suggest that initiation takes place within the valley bottoms to promote replication of all the DNA in the early replicating domains.

We also examined whether there was any discernible pattern in the heights of adjacent peaks that might reflect features affecting the time or efficiency with which they became active. Although there was substantial variability in the peak height order, we showed that there was a marked tendency for adjacent peaks to have similar heights. This is consistent with a 'domino'-style model where the presence of one active replicon cluster increases the probability that a neighbouring replicon cluster becomes active[19,30,53–56]. The mechanism that leads to the preference for domino-style activation remains to be determined, but one can speculate that it depends on the chromatin context within which each Replicon Cluster Domain is situated.

The results presented here give one of the first glimpses of how replication might be organised at the sub-megabase level in somatic metazoan cells, potentially integrating at genomic loci previously disparate results obtained by DNA fibre analysis and high-resolution microscopy. Our results show the potential for mapping DNA replication at a very high temporal resolution. Because cell cycle synchrony is gradually decreased as our cohort of synchronised cells progress through S phase, the technique described here can only be used to analyse the early stages of S phase at high temporal resolution. To further explore these possibilities, technical refinements might be required for achieving higher temporal resolution. Isolating cells that are at very precise stages of S phase is difficult, and the use of alternative possibilities that do not require cell cycle synchronisation might be a better approach. Although we made attempts to limit it, our use of mimosine does create some double-strand breaks, which might reduce the total number of initiation events occurring at later stages. One alternative technical approach would be to sort cells using very fine differences in DNA content. Another possible approach would be to reconstruct the timing programme from the analysis of sites of EdU incorporation in individual S phase cells[7,28,29,57,59,60].

However, despite some drawbacks that are associated with the cell cycle synchrony we have used, this paper reports data that is unbiased and does not depend on assumptions about the structure of S phase. This high-resolution mapping of ours provides new insights into the dynamics and organisation of the genome for duplication.

Conclusions

We show that early replicating DNA is organised into Replicon Cluster Domains that behave as expected of replicon clusters observed by DNA fibre analysis. The domains have a range of sizes that cluster around 500 Mbp. The coordinated activation of different Replicon Cluster Domains can generate the replication timing programme by which the genome is duplicated. Different Replicon Cluster Domains show marked differences in their labelling intensity and we provide evidence that this is at least in part caused by Replicon Cluster Domains being activated at different times in different cells in the population. We also provide evidence that adjacent clusters were preferentially activated in sequence across a group, consistent with the 'domino' model of replication focus activation observed by microscopy.

Data availability

Underlying data

BioStudies: Underlying data for 'The location and development of Replicon Cluster Domains in early replicating DNA'. https://www.ebi.ac.uk/biostudies/studies/S-BSST966[51]

This project contains the following underlying data:

- E1_1_1.fastq.gz (Control, 0-40 min, treatment, replicate 1)
- E1_1_2.fastq.gz (Control, 40-70 min, treatment, replicate 1)
- E1_1_3.fastq.gz (Control, 70-100 min, treatment, replicate 1)
- E1_1_4.fastq.gz (Control, 100-130 min, treatment, replicate 1)
- E1_1_5.fastq.gz (Control, 0-130 min, treatment, replicate 1)
- E1_PD_1.fastq.gz (Pulldown, 0-40 min, treatment, replicate 1)
- E1_PD_2.fastq.gz (Pulldown, 40-70 min, treatment, replicate 1)
- E1_PD_3.fastq.gz (Pulldown, 70-100 min, treatment, replicate 1)
- E1_PD_4.fastq.gz (Pulldown, 100-130 min, treatment, replicate 1)
- E1_PD_5.fastq.gz (Pulldown, 0-130 min, treatment, replicate 1)
- E2_1_1.fastq.gz (Control, 10-40 min, treatment, replicate 2)
- E2_1_10.fastq.gz (Control, 0-130 min, treatment, replicate 3)
- E2_1_11.fastq.gz (Control, 10-40 min, treatment, replicate 4)
- E2_1_12.fastq.gz (Control, 10-40 min, treatment, replicate 5)
- E2_1_13.fastq.gz (Control, 40-70 min, treatment with hydroxyurea, replicate 2)
- E2_1_14.fastq.gz (Control, 70-100 min, treatment with hydroxyurea, replicate 3)
- E2_1_15.fastq.gz (Control, 0-30 min, asynchronous, replicate 4)
- E2_1_17.fastq.gz (Control, 0-60 min, asynchronous unsorted, replicate 1)
- E2_1_2.fastq.gz (Control, 40-70 min, treatment, replicate 2)
- E2_1_20.fastq.gz (Control, 0-40 min, treatment, replicate 1)
- E2_1_21.fastq.gz (Control, 0-40 min, treatment, replicate 2)
- E2_1_22.fastq.gz (Control, 0-40 min, treatment, replicate 3)
- E2_1_23.fastq.gz (Control, 0-30 min, asynchronous, replicate 1)
- E2_1_24.fastq.gz (Control, 0-30 min, asynchronous, replicate 2)
- E2_1_25.fastq.gz (Control, 0-30 min, asynchronous, replicate 3)
- E2_1_26.fastq.gz (Control, 10-40 min, treatment with hydroxyurea, replicate 1)
- E2_1_3.fastq.gz (Control, 70-100 min, treatment, replicate 2)
- E2_1_4.fastq.gz (Control, 100-130 min, treatment, replicate 2)
- E2_1_5.fastq.gz (Control, 0-130 min, treatment, replicate 2)
- E2_1_6.fastq.gz (Control, 10-40 min, treatment, replicate 3)
- E2_1_7.fastq.gz (Control, 40-70 min, treatment, replicate 3)
- E2_1_8.fastq.gz (Control, 70-100 min, treatment, replicate 3)
- E2_1_9.fastq.gz (Control, 100-130 min, treatment, replicate 3)
- E2_PD_1.fastq.gz (Pulldown, 10-40 min, treatment, replicate 2)
- E2_PD_10.fastq.gz (Pulldown, 0-130 min, treatment, replicate 3)
- E2_PD_11.fastq.gz (Pulldown, 10-40 min, treatment, replicate 4)
- E2_PD_12.fastq.gz (Pulldown, 10-40 min, treatment, replicate 5)
- E2_PD_13.fastq.gz (Pulldown, 40-70 min, treatment with hydroxyurea, replicate 2)
- E2_PD_14.fastq.gz (Pulldown, 70-100 min, treatment with hydroxyurea, replicate 3)
- E2_PD_15.fastq.gz (Pulldown, 0-30 min, asynchronous, replicate 4)
- E2_PD_17.fastq.gz (Pulldown, 0-60 min, asynchronous unsorted, replicate 1)
- E2_PD_2.fastq.gz (Pulldown, 40-70 min, treatment, replicate 2)
- E2_PD_20.fastq.gz (Pulldown, 0-40 min, treatment, replicate 1)
- E2_PD_21.fastq.gz (Pulldown, 0-40 min, treatment, replicate 2)
- E2_PD_22.fastq.gz (Pulldown, 0-40 min, treatment, replicate 3)
- E2_PD_23.fastq.gz (Pulldown, 0-30 min, asynchronous, replicate 1)
- E2_PD_24.fastq.gz (Pulldown, 0-30 min, asynchronous, replicate 2)
- E2_PD_25.fastq.gz (Pulldown, 0-30 min, asynchronous, replicate 3)
- E2_PD_26.fastq.gz (Pulldown, 10-40 min, treatment with hydroxyurea, replicate 1)
- E2_PD_3.fastq.gz (Pulldown, 70-100 min, treatment, replicate 2)
- E2_PD_4.fastq.gz (Pulldown, 100-130 min, treatment, replicate 2)

- E2_PD_5.fastq.gz (Pulldown, 0-130 min, treatment, replicate 2)
- E2_PD_6.fastq.gz (Pulldown, 10-40 min, treatment, replicate 3)
- E2_PD_7.fastq.gz (Pulldown, 40-70 min, treatment, replicate 3)
- E2_PD_8.fastq.gz (Pulldown, 70-100 min, treatment, replicate 3)
- E2_PD_9.fastq.gz (Pulldown, 100-130 min, treatment, replicate 3)
- Supp Figure S1.pdf (Supplemental Figure S1)
- Supp Figure S2.pdf (Supplemental Figure S2)
- Supp Figure S3.pdf (Supplemental Figure S3)
- Supp Figure S4.pdf (Supplemental Figure S4)
- Supp Figure S5 v3.pdf (Supplemental Figure S5)
- RT_U2OS_Bone.txt (U2OS timing domain data)

Accession numbers

4D Nucleome: Late fraction S-phase Repliseq on U2OS Tier 2 cells [Homo sapiens]. Accession number 4DNES99LXRYK; https://identifiers.org/4dn:4DNES99LXRYK

4D Nucleome: Early fraction S-phase Repliseq on U2OSTier 2 cells [Homo sapiens]. Accession number 4DNES1P18J2X; https://identifiers.org/4dn:4DNES1P18J2X

Software availability

Source code available from: https://github.com/bartongroup/MG_EarlyReplication

Archived source code at time of publication: https://www.doi.org/10.5281/zenodo.7639072[41]

References

1. Jain M, Koren S, Miga KH, *et al.*: **Nanopore sequencing and assembly of a human genome with ultra-long reads.** *Nat Biotechnol.* 2018; **36**(4): 338–345.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Bellush JM, Whitehouse I: **DNA replication through a chromatin environment.** *Philos Trans R Soc Lond B Biol Sci.* 2017; **372**(1731): 20160287.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Maya-Mendoza A, Tang CW, Pombo A, *et al.*: **Mechanisms regulating S phase progression in mammalian cells.** *Front Biosci (Landmark Ed).* 2009; **14**(11): 4199–4213.
PubMed Abstract | Publisher Full Text
4. Chagin VO, Casas-Delucchi CS, Reinhart M, *et al.*: **4D Visualization of replication foci in mammalian cells corresponding to individual replicons.** *Nat Commun.* 2016; **7**: 11231.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Blow JJ, Ge XQ, Jackson DA: **How dormant origins promote complete genome replication.** *Trends Biochem Sci.* 2011; **36**(8): 405–414.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Moreno A, Carrington JT, Albergante L, *et al.*: **Unreplicated DNA remaining from unperturbed S phases passes through mitosis for resolution in daughter cells.** *Proc Natl Acad Sci U S A.* 2016; **113**(39): E5757–5764.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Wang W, Klein KN, Proesmans K, *et al.*: **Genome-wide mapping of human DNA replication by optical replication mapping supports a stochastic model of eukaryotic replication.** *Mol Cell.* 2021; **81**(14): 2975–2988 e2976.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Jackson DA, Pombo A: **Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells.** *J Cell Biol.* 1998; **140**(6): 1285–1295.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Blow JJ, Ge XQ: **A model for DNA replication showing how dormant origins safeguard against replication fork failure.** *EMBO Rep.* 2009; **10**(4): 406–412.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Takebayashi SI, Sugimura K, Saito T, *et al.*: **Regulation of replication at the R/G chromosomal band boundary and pericentromeric heterochromatin of mammalian cells.** *Exp Cell Res.* 2005; **304**(1): 162–174.
PubMed Abstract | Publisher Full Text
11. Parplys AC, Seelbach JI, Becker S, *et al.*: **High levels of RAD51 perturb DNA replication elongation and cause unscheduled origin firing due to impaired CHK1 activation.** *Cell Cycle.* 2015; **14**(19); 3190–3202.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Patel PK, Arcangioli B, Baker SP, *et al.*: **DNA replication origins fire stochastically in fission yeast.** *Mol Biol Cell.* 2006; **17**(1): 308–316.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Rhind N, Yang SCH, Bechhoefer J: **Reconciling stochastic origin firing with defined replication timing.** *Chromosome Res.* 2010; **18**(1): 35–43.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Guilbaud G, Murat P, Wilkes HS, *et al.*: **Determination of human DNA replication origin position and efficiency reveals principles of initiation zone organisation.** *Nucleic Acids Res.* 2022; **50**(13): 7436–7450.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Ma H, Samarabandu J, Devdhar RS, *et al.*: **Spatial and temporal dynamics of DNA replication sites in mammalian cells.** *J Cell Biol.* 1998; **143**(6): 1415–1425.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Nakayasu H, Berezney R: **Mapping replicational sites in the eucaryotic cell nucleus.** *J Cell Biol.* 1989; **108**(1): 1–11.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Sparvoli E, Levi M, Rossi E: **Replicon clusters may form structurally stable complexes of chromatin and chromosomes.** *J Cell Sci.* 1994; **107**(Pt 11): 3097–3103.
PubMed Abstract | Publisher Full Text
18. Ferreira J, Paolella G, Ramos C, *et al.*: **Spatial organization of large-scale chromatin domains in the nucleus: a magnified view of single chromosome territories.** *J Cell Biol.* 1997; **139**(7): 1597–1610.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Farkash-Amar S, Lipson D, Polten A, *et al.*: **Global organization of replication time zones of the mouse genome.** *Genome Res.* 2008; **18**(10): 1562–1570.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Yamazaki S, Ishii A, Kanoh Y, *et al.*: **Rif1 regulates the replication timing domains on the human genome.** *EMBO J.* 2012; **31**(18): 3667–3677.
PubMed Abstract | Publisher Full Text | Free Full Text
21. Takahashi S, Miura H, Shibata T, *et al.*: **Genome-wide stability of the DNA replication program in single mammalian cells.** *Nat Genet.* 2019; **51**(3): 529–540.
PubMed Abstract | Publisher Full Text
22. Hiratani I, Ryba T, Itoh M, *et al.*: **Global reorganization of replication domains during embryonic stem cell differentiation.** *PLoS Biol.* 2008; **6**(10): e245.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Ryba T, Hiratani I, Lu J: **Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types.** *Genome Res.* 2010; **20**(6): 761–770.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Marchal C, Sima J, Gilbert DM: **Control of DNA replication timing in the 3D genome.** *Nat Rev Mol Cell Biol.* 2019; **20**(12): 721–737.
PubMed Abstract | Publisher Full Text

25. Hiratani I, Ryba T, Itoh M, *et al.*: **Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis.** *Genome Res.* 2010; **20**(2): 155–169.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Rivera-Mulia JC, Buckley Q, Sasaki T, *et al.*: **Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells.** *Genome Res.* 2015; **25**(8): 1091–1103.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Hadjadj D, Denecker T, Maric C, *et al.*: **Characterization of the replication timing program of 6 human model cell lines.** *Genom Data.* 2016; **9**: 113–117.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Dileep V, Gilbert DM: **Single-cell replication profiling to measure stochastic variation in mammalian replication timing.** *Nat Commun.* 2018; **9**(1): 427.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Zhao PA, Sasaki T, Gilbert DM: **High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells.** *Genome Biol.* 2020; **21**(1): 76.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Gillespie PJ, Blow JJ: **Clusters, factories and domains, The complex structure of S-phase comes into focus.** *Cell Cycle.* 2010; **9**(16): 3218–3226.
PubMed Abstract | Publisher Full Text | Free Full Text

31. O'Keefe RT, Henderson SC, Spector DL: **Dynamic organization of DNA replication in mammalian cell nuclei, spatially and temporally defined replication of chromosome-specific alpha-satellite DNA sequences.** *J Cell Biol.* 1992; **116**(5): 1095–1110.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Dimitrova DS, Gilbert DM: **The spatial position and replication timing of chromosomal domains are both established in early G1 phase.** *Mol Cell.* 1999; **4**(6): 983–993.
PubMed Abstract | Publisher Full Text

33. Leonhardt H, Rahn HP, Weinzierl P, *et al.*: **Dynamics of DNA replication factories in living cells.** *J Cell Biol.* 2000; **149**(2): 271–280.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Ge XQ, Blow JJ: **Chk1 inhibits replication factory activation but allows dormant origin firing in existing factories.** *J Cell Biol.* 2010; **191**(7): 1285–1297.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Berezney R, Dubey DD, Huberman JA: **Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci.** *Chromosoma.* 2000; **108**(8): 471–484.
PubMed Abstract | Publisher Full Text

36. Nakamura H, Morita T, Sato C: **Structural organizations of replicon domains during DNA synthetic phase in the mammalian nucleus.** *Exp Cell Res.* 1986; **165**(2): 291–297.
PubMed Abstract | Publisher Full Text

37. Ryba T, Battaglia D, Pope BD, *et al.*: **Genome-scale analysis of replication timing, from bench to bioinformatics.** *Nat Protoc.* 2011; **6**(6): 870–895.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods.* 2012; **9**(4): 357–359.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools.** *Gigascience.* 2021; **10**(2): giab008.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Barnett DW, Garrison EK, Quinlan AR, *et al.*: **BamTools: a C++ API and toolkit for analyzing and managing BAM files.** *Bioinformatics.* 2011; **27**(12): 1691–1692.
PubMed Abstract | Publisher Full Text | Free Full Text

41. da Costa-Nunes JA, Gierlinski M, Sasaki T, *et al.*: **The location and development of Replicon Cluster Domains in early replicating DNA.** *GitHub.* 2023.
https://github.com/bartongroup/MG_EarlyReplication

42. Ji H, Jiang H, Man W, *et al.*: **An integrated software system for analyzing ChIP-chip and ChIP-seq data.** *Nat Biotechnol.* 2008; **26**(11): 1293–1300.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Liang K, Keles S: **Normalization of ChIP-seq data with control.** *BMC Bioinformatics.* 2012; **13**: 199.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Mosca PJ, Dijkwel PA, Hamlin JL: **The plant amino acid mimosine may inhibit initiation at origins of replication in Chinese hamster cells.** *Mol Cell Biol.* 1992; **12**(19): 4375–4383.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Krude T: **Mimosine arrests proliferating human cells before onset of DNA replication in a dose-dependent manner.** *Exp Cell Res.* 1999; **247**(1): 148–159.
PubMed Abstract | Publisher Full Text

46. Galgano PJ, Schildkraut CL: **G1/S phase synchronization using mimosine arrest.** *CSH Protoc.* 2006; **2006**(4): pdb.prot4488.
PubMed Abstract | Publisher Full Text

47. Ji C, Marnett LJ, Pietenpol JA: **Cell cycle re-entry following chemically-induced cell cycle synchronization leads to elevated p53 and p21 protein levels.** *Oncogene.* 1997; **15**(22): 2749–2753.
PubMed Abstract | Publisher Full Text

48. Wang G, Miskimins R, Miskimins WK: **Mimosine arrests cells in G1 by enhancing the levels of p27^{Kip1}.** *Exp Cell Res.* 2000; **254**(1): 64–71.
PubMed Abstract | Publisher Full Text

49. Szuts D, Krude T: **Cell cycle arrest at the initiation step of human chromosomal DNA replication causes DNA damage.** *J Cell Sci.* 2004; **117**(Pt 21): 4897–4908.
PubMed Abstract | Publisher Full Text

50. Park SY, Im JS, Park SR, *et al.*: **Mimosine arrests the cell cycle prior to the onset of DNA replication by preventing the binding of human Ctf4/And-1 to chromatin via Hif-1alpha activation in *HeLa* cells.** *Cell Cycle.* 2012; **11**(4): 761–766.
PubMed Abstract | Publisher Full Text

51. da Costa-Nunes JA, Gierlinski M, Sasaki T, *et al.*: **The location and development of Replicon Cluster Domains in early replicating DNA Supp Figures S1-S5.** *Biostudies.* 2023.
https://www.ebi.ac.uk/biostudies/studies/S-BSST966

52. Gilbert DM, Neilson A, Miyazawa H, *et al.*: **Mimosine arrests DNA synthesis at replication forks by inhibiting deoxyribonucleotide metabolism.** *J Biol Chem.* 1995; **270**(16): 9597–9606.
PubMed Abstract | Publisher Full Text

53. Sporbert A, Gahl A, Ankerhold R, *et al.*: **DNA polymerase clamp shows little turnover at established replication sites but sequential de novo assembly at adjacent origin clusters.** *Mol Cell.* 2002; **10**(6): 1355–1365.
PubMed Abstract | Publisher Full Text

54. Sadoni N, Cardoso MC, Stelzer EH, *et al.*: **Stable chromosomal units determine the spatial and temporal organization of DNA replication.** *J Cell Sci.* 2004; **117**(Pt 22): 5353–5365.
PubMed Abstract | Publisher Full Text

55. Maya-Mendoza A, Olivares-Chauvet P, Shaw A, *et al.*: **S phase progression in human cells is dictated by the genetic continuity of DNA foci.** *PLoS Genet.* 2010; **6**(4): e1000900.
PubMed Abstract | Publisher Full Text | Free Full Text

56. Lob D, Lengert N, Chagin VO, *et al.*: **3D replicon distributions arise from stochastic initiation and domino-like DNA replication progression.** *Nat Commun.* 2016; **7**: 11207.
PubMed Abstract | Publisher Full Text | Free Full Text

57. Miura H, Takahashi S, Shibata T, *et al.*: **Mapping replication timing domains genome wide in single mammalian cells with single-cell DNA replication sequencing.** *Nat Protoc.* 2020; **15**(12): 4058–4100.
PubMed Abstract | Publisher Full Text

58. Pope BD, Ryba T, Dileep V, *et al.*: **Topologically associating domains are stable units of replication-timing regulation.** *Nature.* 2014; **515**(7527): 402–405.
PubMed Abstract | Publisher Full Text | Free Full Text

59. Klein KN, Zhao PA, Lyu X, *et al.*: **Replication timing maintains the global epigenetic state in human cells.** *Science.* 2021; **372**(6540): 371–378.
PubMed Abstract | Publisher Full Text | Free Full Text

60. Bartlett DA, Dileep V, Baslan T, *et al.*: **Mapping Replication Timing in Single Mammalian Cells.** *Curr Protoc.* 2022; **2**(1): e334.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 2

Reviewer Report 29 September 2023

https://doi.org/10.21956/wellcomeopenres.22056.r65746

✔ **Zhongqing Ren**
Indiana University Bloomington, Bloomington, Indiana, USA

The authors have addressed my concerns with the original manuscript.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Chromosome organization and segregation

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 07 September 2023

https://doi.org/10.21956/wellcomeopenres.22056.r65747

✔ **Itamar Simon**
Department of Microbiology and Molecular Genetics, IMRIC, Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel
Avraham Greenberg
Department of Microbiology and Molecular Genetics, IMRIC, Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

In light of the authors' answers, and the changes that they have made, I would give the revised article the status of 'approved', and I thank the authors for addressing my concerns.

Competing Interests: No competing interests were disclosed.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 07 September 2023

https://doi.org/10.21956/wellcomeopenres.22056.r65748

Ichiro Hiratani
RIKEN Center for Biosystems Dynamics Research, Kobe, Japan

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 18 July 2023

https://doi.org/10.21956/wellcomeopenres.20781.r60829

Zhongqing Ren
Indiana University Bloomington, Bloomington, Indiana, USA

In this study, the authors employed EdU pull-down combined with high-throughput sequencing on synchronized human cells to study the replication cluster domains in early S phase. This whole-genome scale method complements DNA fiber analysis, which analyzes individual molecules.

The results are consistent with the previous studies and revealed more details at fine-scale about the early replication cluster domains. Taking advantage of the higher resolution of the data, the authors defined the sizes, spacing and expansion rates of the replication cluster domains.

Furthermore, the authors provided evidence supporting that late-activated replication cluster

domains could fill the "gaps" between the early-activated replication cluster domains, and a "domino" activation model where the early-fired replication cluster domains increase the probability of the activation of the adjacent replication cluster domains. Overall, the experiments were well controlled and the data were nicely analyzed. Although the cells were not well synchronized, the article still provides valuable and interesting information. I do not request more experiments but have a few comments:

Major:

1. From Figure S1, one can see that the cells were not synchronized very well. Even after FCAS sorting and selection, some chromosome regions have higher copy number than the other (Figure S2), suggesting some regions may have already started replication. The authors should consider the effects of the pre-fired origins when analyze, interpret and discuss the results. For the future experiments, the authors may need to optimize their synchronization protocol (e.g., double thymidine block (Chen et al., 2018 [1])) or use a different synchronization method (e.g., CDKs inhibition).

Minor:

1. "*This increase in DNA content can be seen in synchronised cells pulsed continuously with EdU for 130 min (Figure 1b, 0–130 min).*" It is hard to see the increase in whole DNA content in Fig 1b. It would be helpful to provide some quantified results.

2. It would be helpful if the authors can explain why the Ricker wavelet is used for analyzing this type of data.

3. The color scales for the heatmaps are absent in Figure 2C and 3. The color contrast is not great. May want to consider to change the color scheme.

4. "*This is in line with the two later expansion rates*". The expansion between the 2nd and 3rd time points is not really in line with 60-90kb. There is an increase of expansion rate for these four time points, which is better to be mentioned and discussed.

5. "*The median separation between Replicon Cluster Domains is ~1.2 Mbp (Figure 5) and if they have a width of ~500 kb in size flanking forks progressing ~360 kb over the two-hour time course only extend the median cluster to a width of ~860 kb, which is two thirds of the distance required. This suggests that complete replication of valley DNA will depend on further initiation events.*". This argument is not strong. The separated replication cluster domains might fill the entire valley after three hours. Additional information is needed to justify the statement, e.g., the replication forks usually stay on the chromosome for xx hours.

6. "*There is also good evidence for initiation in the valleys between Replicon Cluster Domains from the replication data*". It is hard to find the related region in these figures. It will be helpful to use arrows to point out the related regions.

7. "*This would also be consistent with the considerable variation in the height of the peaks representing the Replicon Cluster Domains which could be caused by them becoming active at different times in different cells in the population.*" My understanding is that replication cluster domains with the lower height of peaks generally activate later than those with higher peaks in the population. It that correct? If so, the sentence quoted above is not accurate and need to be revised.

8. For Figure S2, more details about how each control was processed are needed. What is the difference between 1% and 99% samples of the G1 FACS-sorted cells?

References

1. Chen G, Deng X: Cell Synchronization by Double Thymidine Block.*Bio Protoc*. 2018; **8** (17).
PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Chromosome organization and segregation

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 14 Aug 2023

J Julian Blow

Major:

1. *From Figure S1, one can see that the cells were not synchronized very well. Even after FCAS sorting and selection, some chromosome regions have higher copy number than the other (Figure S2), suggesting some regions may have already started replication. The authors should consider the effects of the pre-fired origins when analyze, interpret and discuss the results. For the future experiments, the authors may need to optimize their synchronization protocol (e.g., double thymidine block (Chen et al., 2018 [1])) or use a different synchronization method (e.g., CDKs inhibition).*

Achieving the best possible synchronisation procedure was one of the major challenges of this work. Our protocol allowed a significantly tighter burst of S phase entry compared to standard methods such as double thymidine synchronisation. Despite our best efforts, entry into S phase was staggered over a period of about 45 minutes as shown in Figures 1, S1 and S2. There is little evidence to suggest that cells leaked through the mimosine block as suggested by the reviewer. Figure S2 shows increasing pulses of EdU incorporation after mimosine release (0, 2.5, 5, 10, 15, 20, 25, 30 and 40 minutes after mimosine release). Virtually no EdU incorporation is seen for the first 10 minutes after mimosine release and provides evidence that the vast majority of cells incorporating EdU do so from a non-replicating G1 DNA content. There are a relatively small number of cells that incorporate EdU with a mid- to late- S phase DNA content, indicative of cells that didn't complete the previous S phase after thymidine release. Almost all of these cells were lost following the FACS sort for cells with a near-G1 DNA content. For a traditional block-and-release synchronisation procedure we think that these cells were synchronised very well. However, we agree that alternative technical approaches, such as are mentioned at the end of the paper, would improve the synchrony still further.

Minor:

1. *"This increase in DNA content can be seen in synchronised cells pulsed continuously with EdU for 130 min (Figure 1b, 0–130 min)." It is hard to see the increase in whole DNA content in Fig 1b. It would be helpful to provide some quantified results.*

   For all pulses where EdU incorporation is evident in Figure 1 and Figure S1 (i.e. pulses longer than 15 min), the EdU-labelled DNA very clearly moves to the right on the x-axis away from the peak of G1 cells, representing an increase in total DNA content. To make this more obvious we have replotted the FACS data in Figure 1 and Supplementary Figure S1 to decrease the extent of the y axis (DNA content) thereby making the increase in DNA content more obvious. We also have amended the above text in the Results section to describe this: "*This increase in DNA content can be seen in synchronised cells pulsed with EdU for 100-130 min (Figure 1b), where the total increase in DNA content is ~15%; shorter pulses of EdU (Underlying data: Supplementary Figure S1) also show increases in DNA content, with ~6% increase in DNA content after 40 min. This slightly slowed progression through S phase could be due to mimosine inducing a small number of double strand breaks (47-50) or reducing cellular dNTP pools (52).*" Please also see response to Reviewer 2's second major point.

2. *It would be helpful if the authors can explain why the Ricker wavelet is used for analyzing this type of data.*

   The Ricker wavelet is one of the most basic wavelets used for analysis and is more likely to approximate our data than other more complex wavelets. We have added the following sentence to the Results section when we introduce the wavelet analysis: "*We chose to use a Ricker wavelet (Supplementary Figure S4) for the analysis because it is a simple symmetrical wavelet that is derived from a Gaussian distribution and therefore makes minimal assumptions about the expected shape of the peaks.*"

3. *The color scales for the heatmaps are absent in Figure 2C and 3. The color contrast is not great. May want to consider to change the color scheme.*

   We have added a heatmap colour scheme to the Figure and added a sentence to the legend saying: “Dark blue represents a signal of zero or below, and red represents a signal of one.” The relative lack of orange and red in Figure 2 is because the 500 kb peaks are extremely small at this scale; they only become apparent after zooming in to smaller regions as in Figure 3.

4. *"This is in line with the two later expansion rates". The expansion between the 2nd and 3rd time points is not really in line with 60-90kb. There is an increase of expansion rate for these four time points, which is better to be mentioned and discussed.*

   As requested by Reviewer 2, we have added some more statistical information to this figure (panel 7G). We have also recalculated the rate information in panels 7E and 7F in kb/min for better comparison with 7G. This shows that between the 2nd and 3rd time points, the width increases by 1.3 kb/min (wavelet analysis) and 0.37 kb/min (Gaussian analysis), which we agree is on the low side for two forks. We have therefore changed the phrase to "*This is in line with the later expansion rates*".

5. *"The median separation between Replicon Cluster Domains is ~1.2 Mbp (Figure 5) and if they have a width of ~500 kb in size flanking forks progressing ~360 kb over the two-hour time course only extend the median cluster to a width of ~860 kb, which is two thirds of the distance required. This suggests that complete replication of valley DNA will depend on further initiation events.". This argument is not strong. The separated replication cluster domains might fill the entire valley after three hours. Additional information is needed to justify the statement, e.g., the replication forks usually stay on the chromosome for xx hours.*

   The reviewer is correct that the assumption of the sentence was about replication within the period of the analysis. For greater clarity we have therefore changed the sentence to: “This suggests that complete replication of valley DNA within a two hour period would depend on further initiation events.”

6. *"There is also good evidence for initiation in the valleys between Replicon Cluster Domains from the replication data". It is hard to find the related region in these figures. It will be helpful to use arrows to point out the related regions.*

   The examples in Figure 8 are worked through, with the brown bars showing expected fork progression. In case this isn’t clear enough we have amended the sentence describing this (as underlined) to: “*However, many gaps <u>between the brown bars</u> still remain in the last time point so it is clear that this cannot account for replication of all the valley DNA*.”

7. *"This would also be consistent with the considerable variation in the height of the peaks representing the Replicon Cluster Domains which could be caused by them becoming active at different times in different cells in the population." My understanding is that*

replication cluster domains with the lower height of peaks generally activate later than those with higher peaks in the population. It that correct? If so, the sentence quoted above is not accurate and need to be revised.

Yes the reviewer's interpretation corresponds to what we were trying to convey, though the text was perhaps not so clear. To clarify, we have amended the sentence to: "This would also be consistent with the considerable variation in the height of the peaks representing the Replicon Cluster Domains ... which could be caused by them becoming active at different times, with lower peak heights representing Domains that tend to become active at later times."

8. *For Figure S2, more details about how each control was processed are needed. What is the difference between 1% and 99% samples of the G1 FACS-sorted cells?*

   The processing and normalisation of the data is described in quite some detail in Methods under the headings "Se*quencing data processing" and "Background subtraction and normalisation".* For the difference between the 1% and 99% data (which was unexpected), we have added the following sentence to the legend to Figure S2 to explain why we think the last three samples (including the G1_99 sample) are much less uniform than the others: "We believe that the more erratic profiles of the last three samples is due to the fact that considerably more DNA was submitted for library preparation, which resulted in competition between DNA and limiting amounts of primer." We did not use these last three data sets for normalisation.

Competing Interests: No competing interests were disclosed.

Reviewer Report 03 May 2023

https://doi.org/10.21956/wellcomeopenres.20781.r56116

?

Itamar Simon
Department of Microbiology and Molecular Genetics, IMRIC, Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel
Avraham Greenberg
Department of Microbiology and Molecular Genetics, IMRIC, Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

In this paper the authors used a combination of synchronization, EdU pulses and pulldown, and careful normalization and wavelet analysis in order to visualize the replication of origin clusters at a genomic level in early S phase. The detailed and in-depth bioinformatics in this paper managed

to address a few aspects of replication organization, which have previously mainly been addressed by DNA fiber analyses. This includes the observations that:

1. replication domains are built of several replicons that fire simultaneously, which was previously deduced by a combination of replication domain sizes (genomics) and replicon sizes (DNA combing);

2. replication domains extend in both directions by single forks moving in each direction at the ends of the domains (this is the common interpretation of the TTR that usually flanks replication domains);

3. expansion of domains by TTR is not sufficient to fill the gaps between early domains thus one must assume the initiation of later domains (this is the common interpretation of replication maps);

4. early domains somehow promote the activation of adjacent later domains (the domino model that was previously suggested was based on microscopy and DNA fiber analyses).

Overall, the work is an elegant application of a number of analytic techniques to biological questions, which were previously studied using mainly DNA combing techniques. Thus, the findings in this paper are not new, but provide new support for previously published findings using a different type of technology.

Major points:

- The authors have not demonstrated why EdU immune-precipitation in early S phase cells provides higher resolution than Repli-seq of non-synchronized cells. Would similar wavelet analysis of Repli-seq data not also show similar trends? If the aim is to demonstrate that domains are replicon clusters, couldn't this be done by analyzing Repli-seq data without the need to synchronize the cells and pulldown with EdU?

- We are surprised not to see substantial progress along the cell cycle (i.e., PI intensity) even after 2 hours. Did the authors take any timepoints (for **Figure 1B/S1**) later than 2 hours? Could the authors add quantification (for example using gates) to show the progress along the x-axis in **Figure 1B**? The lack of progression raises a question about how representative the results are; do they reflect a normal unperturbed cell cycle or do they represent a specific case in which the entry into S phase is very slow.

Minor points:

1. In the second paragraph of the Introduction the authors state "*Indeed, of the total number of potentially available replication origins present in diploid mammalian cells – roughly 100,000,000 – only approximately 10% will be used to initiate DNA replication*". Is the 100,000,000 a typo? We are more familiar with lower estimates, specifically 500,000 licensed origins.

2. The source for the RT data is provided as https://www2.replicationdomain.com/database.php, but this link is not accessible. As such the data is currently not accessible to the public.

3. The inset in **Figure 1B** is not explained.

4. **Figure 2C** - scale of graph missing (and explanation of the units of the score)

5. **Reference 48** seems the wrong reference for page 9 ("*This is consistent with recent data on single DNA fibres which shows the high degree of stochasticity that occurs in origin firing*").

6. **Figure 4A** and **B** – What does the 0-40 timepoint add over the 10-40 timepoint? Also why is the 10kb and the 50kb comparison important? Also, 4c and 4d come to justify the 70 percentile cutoff, but this has already been relied on in **Figure 3**. Perhaps the figure order could be swapped between **Figures 3** and **4**? Or **4** could be moved to supplementary.

7. When discussing **Figure 4D**, page 12, the authors say that the variance between the clusters is due to stochasticity between cells. We can think of 2 alternative explanations. Based on what we have observed in published data, and in our own lab's unpublished data, inter-origin distance has as much variance as the 20-fold variance shown in **Figure 4D**. Alternatively, the variance could reflect locus to locus variation in the time that different Replicon Cluster Domains become active. Although these three options are discussed in the discussion it might help to mention all three in the results as well, as the conclusion that stochasticity between cells is responsible remains not fully proven.

8. **Figure 5B** – how the early replication domain are defined?

9. Regarding the Isolated Replicon Cluster Domains:

    1. **Figures 7C-F** are based on only 72 isolated domains. How sure are they that the conclusions derived from the isolated domains are relevant to more crowded regions?

    2. The authors say "*We rejected from further analysis 51 peaks*" was there a biological justification for removing these, instead of expanding the range of widths (e.g, 100-1500).

10. Statistical tests are missing in **figures 7E-F, 9B-C** and **10**.

11. In my understanding the edges of the timing domains reflect what is generally referred to as TTR. The authors should mention this, for example when discussing the valley filling in **Figure 8**, or in the discussion "*As cells progress through S phase, the width of these peaks grows at rates consistent with the progression of a single pair of replication forks at the extreme edges of each peak*.".

12. In **Figure 8**, the expansion of the bars should start from the 40-70m timepoint in line with what the authors wrote on page 14 ("*The lack of significant growth between the first two time points …*").

13. Near the end of the results the authors say "*the mean spacing is 817 kb"*. This appears to be a mistake, based on **Figure 5** (0.817 is the number of peaks per Mbp. The real number should be 1.2 Mb spacing between peaks). The rest of the analysis of the adjacent Replicon Cluster Domains for **Figure 10** should be recalculated accordingly.

14. **Figures 10C** and **D** use an elegant metric but do not accurately measure the domino effect,

as they assume that the outermost origin must be the first one to be activated. For example, by this logic, the 2-1-3 order from **Figure 10A** would not receive the maximal score, despite its appearing like a clear case of domino progression. Regarding **10A and B** it would be clearer/more persuasive if the authors combined all of the interrupted progression/uninterrupted progression and compared those prevalences. For example, in 10A, display the prevalence of 1-2-3 and 2-1-3, compared to the interrupted pattern of 1-3-2. And in 10B, compare the uninterrupted 1-2-3-4, 2-1-3-4, 3-1-2-4, 4-1-2-3 with the others.

15. Typo in conclusion, page 19 first line – 'part part'.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: DNA replication timing, genomics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 14 Aug 2023

J Julian Blow

Major points:

- *The authors have not demonstrated why EdU immune-precipitation in early S phase cells provides higher resolution than Repli-seq of non-synchronized cells. Would similar wavelet analysis of Repli-seq data not also show similar trends? If the aim is to demonstrate that domains are replicon clusters, couldn't this be done by analyzing Repli-seq data without the need to synchronize the cells and pulldown with EdU?*

The issue is a matter of the high degree of synchronisation required. Repli-seq provides only a snapshot of populations in early or late S phase, while our release from G1/S synchronization and detection of EdU incorporating regions at various time points gives us much more detailed dynamics. As shown in Figure 2 and Supplementary Figure S2 our data is highly consistent with previously published timing studies but shows the more detailed structure of cells as they progress through early S phase. As mentioned in the discussion, other ways of synchronisation could be used for future work.

- *We are surprised not to see substantial progress along the cell cycle (i.e., PI intensity) even after 2 hours. Did the authors take any timepoints (for **Figure 1B/S1**) later than 2 hours? Could the authors add quantification (for example using gates) to show the progress along the x-axis in **Figure 1B**? The lack of progression raises a question about how representative the results are; do they reflect a normal unperturbed cell cycle or do they represent a specific case in which the entry into S phase is very slow.*

  During the time course, an increase in total cellular DNA content can be observed, but we agree it is less than might be expected. This is presumably a consequence of the synchronisation protocol – it is known that mimosine can cause double strand breaks, so we tried to minimise mimosine exposure. To reflect this, we have added a sentence to the Discussion: "Although we made attempts to limit it, our use of mimosine does create some double-strand breaks, which might reduce the total number of initiation events occurring at later stages." We have also amended the text in the Results section to describe this: "This increase in DNA content can be seen in synchronised cells pulsed with EdU for 100-130 min (Figure 1b), where the total increase in DNA content is ~15%; shorter pulses of EdU (Underlying data: Supplementary Figure S1) also show increases in DNA content, with ~6% increase in DNA content after 40 min. This slightly slowed progression through S phase could be due to mimosine inducing a small number of double strand breaks (47-50) or reducing cellular dNTP pools (52)." Please also see response to Reviewer 3's minor point 1.

Minor points:

1. *In the second paragraph of the Introduction the authors state "Indeed, of the total number of potentially available replication origins present in diploid mammalian cells – roughly 100,000,000 – only approximately 10% will be used to initiate DNA replication". Is the 100,000,000 a typo? We are more familiar with lower estimates, specifically 500,000 licensed origins.*

   This was a consequence of us trying to be too generic and talking about mammalian cells. We have replaced with estimates about human diploid cells, and given the figure of 500,000.

2. *The source for the RT data is provided as https://www2.replicationdomain.com/database.php, but this link is not accessible. As such the data is currently not accessible to the public.*

   We have now uploaded this data to the EBI site (https://ftp.ebi.ac.uk/biostudies/fire/S-

BSST/966/S-BSST966/Files/RT_U2OS_Bone.txt), as the replicationdomain site is not currently operational.

3. *The inset in Figure 1B is not explained.*

   In response to Reviewer 3's minor point 1 we have replotted the FACS graphs in Figure 1 and removed the inset histograms .

4. ***Figure 2C** - scale of graph missing (and explanation of the units of the score)*

   We have added a heatmap colour scheme to the Figure and added a sentence to the legend saying: "Dark blue represents a signal of zero or below, and red represents a signal of one." Both chromosome position on the x axis and wavelet scale on the y axis are labelled.

5. **Reference 48** seems the wrong reference for page 9 ("*This is consistent with recent data on single DNA fibres which shows the high degree of stochasticity that occurs in origin firing*").

   We've now used the correct reference.

6. ***Figure 4A** and **B** – What does the 0-40 timepoint add over the 10-40 timepoint? Also why is the 10kb and the 50kb comparison important? Also, 4c and 4d come to justify the 70 percentile cutoff, but this has already been relied on in **Figure3**. Perhaps the figure order could be swapped between **Figures 3** and **4**? Or **4** could be moved to supplementary.*

   The reason for providing all the different analyses – both 10kb and 50 kb bins, and both 0-40 and 10-40 – is to show that the presence of a peak at around 500kb is quite robust to different data being used and how it is divided up. Since this is a major conclusion of the work, we think it best to show all the data. We think it is easier for readers to see the exemplar data in Fig 3 first, then see the numerical analysis in Figure 4, although following the text does require readers to go back and forth between them.

7. *When discussing **Figure 4D**, page 12, the authors say that the variance between the clusters is due to stochasticity between cells. We can think of 2 alternative explanations. Based on what we have observed in published data, and in our own lab's unpublished data, inter-origin distance has as much variance as the 20-fold variance shown in **Figure 4D**. Alternatively, the variance could reflect locus to locus variation in the time that different Replicon Cluster Domains become active. Although these three options are discussed in the discussion it might help to mention all three in the results as well, as the conclusion that stochasticity between cells is responsible remains not fully proven.*

   The text on page 12 discusses two possible explanations for the differences in peak height: cell-to-cell variation in origin density or cell-to-cell variation in the time that Replicon Cluster Domains become active in different cells. I think this latter explanation is equivalent to the locus to locus variation suggested by the reviewer.

We have amended the text slightly to make this clearer and emphasise that in each case this would be due to cell-to-cell variation: "*The difference in peak heights could be caused by a combination of two different effects: it could represent cell-to-cell variation in the densities of active origins in each Replicon Cluster Domain or it could represent some stochasticity in the time that Replicon Cluster Domains become active.*"

8. ***Figure 5B*** *– how the early replication domain are defined?*

   It was defined from the early timing domain data as illustrated in Figure 2b. The legend to Fig 5b has been amended to indicate this.

9. *Regarding the Isolated Replicon Cluster Domains:*
   1. ***Figures 7C-F*** *are based on only 72 isolated domains. How sure are they that the conclusions derived from the isolated domains are relevant to more crowded regions?*

      We can't be totally sure because in more crowded regions the edges cannot be clearly determined. However, these crowded peaks do seem to grow in a similar manner – for example see Figure 8.

   2. *The authors say "We rejected from further analysis 51 peaks" was there a biological justification for removing these, instead of expanding the range of widths (e.g, 100-1500).*

      This was just because they exceeded the data limits we had set. Visual analysis suggested that these peaks were not being correctly fitted by the wavelet or Gaussian analysis due to either their unusual shape or being affected by other EdU incorporation impinging on the peaks, likely due to additional initiation events. This is likely related to point 3 by Reviewer 1 about timing transition regions and this reviewer's point 11, but our data are not clear enough to unequivocally interpret these in terms of TTRs.

10. *Statistical tests are missing in* ***figures 7E-F, 9B-C*** *and* ***10****.*

    For Figure 7E-F we have added the missing statistical information to the legend. We have also added a new panel 7G which shows that the peak width increases are statistically significant. For Figure 9 we have added a new panel D which shows that the valley filling is statistically significant. For Figure 10 we have added a new panel E which shows that the positive value of the adjacency metric is highly significant. We also proved Standard Errors rather than Standard Deviations for Figures 7 and 9, as this better reflects our confidence in the mean values presented.

11. *In my understanding the edges of the timing domains reflect what is generally referred to as TTR. The authors should mention this, for example when discussing the valley filling in* ***Figure 8****, or in the discussion "As cells progress through S phase, the width of these peaks grows at rates consistent with the progression of a single pair of replication forks at the extreme edges of each peak.".*

The 'valleys' analysed in Figure 8 are not TTRs as previously described, because TTRs are defined as being at the edges of individual Timing Domains, but the valleys we analyse are between two Replicon Cluster Domains that lie within an individual Timing Domain.

12. *In **Figure 8**, the expansion of the bars should start from the 40-70m timepoint in line with what the authors wrote on page 14 ("The lack of significant growth between the first two time points ...").*

    Because we wanted to stress that it is highly likely that further initiation events between the peaks is required, we wanted to show the most extreme example where the forks move outwards straight away. But the reviewer makes a good point that this movement is not clearly evident in the first time point.
    We have added the following sentence to the description of this data: "*In addition, we showed in Figure 7 that fork-driven expansion of the edges of Replicon Cluster Domains is not clearly seen in the first time point due to the continued entry of cells into S phase.*"

13. *Near the end of the results the authors say "the mean spacing is 817 kb". This appears to be a mistake, based on **Figure 5**(0.817 is the number of peaks per Mbp. The real number should be 1.2 Mb spacing between peaks). The rest of the analysis of the adjacent Replicon Cluster Domains for **Figure 10** should be recalculated accordingly.*

    The wording in the text was unclear here and has been replaced by "Within individual timing domains, the mean distance between adjacent Replicon Cluster Domains is 817 kb (Figure 5b)." The rationale for the definition of 'adjacent' (which is somewhat arbitrary anyway) still stands.

14. ***Figures 10C** and **D** use an elegant metric but do not accurately measure the domino effect, as they assume that the outermost origin must be the first one to be activated. For example, by this logic, the 2-1-3 order from **Figure 10A** would not receive the maximal score, despite its appearing like a clear case of domino progression. Regarding **10A and B** it would be clearer/more persuasive if the authors combined all of the interrupted progression/uninterrupted progression and compared those prevalences. For example, in 10A, display the prevalence of 1-2-3 and 2-1-3, compared to the interrupted pattern of 1-3-2. And in 10B, compare the uninterrupted 1-2-3-4, 2-1-3-4, 3-1-2-4, 4-1-2-3 with the others.*

    This is one of the problems with using the 'domino' metaphor, as there is a possibility that the first domino in the stack is internal and should therefore fall in both directions. To indicate this possibility without confusing the reader, we have re-ordered the permutations in 10A and B so that in moving left to right it moves from less domino-like to more domino-like.

15. *Typo in conclusion, page 19 first line – 'part part'.*
    Corrected.

Competing Interests: No competing interests were disclosed.

Reviewer Report 26 April 2023

https://doi.org/10.21956/wellcomeopenres.20781.r56117

?

Ichiro Hiratani
RIKEN Center for Biosystems Dynamics Research, Kobe, Japan

da Costa-Nunes JA et al., The location and development of Replicon Cluster Domains in early replicating DNA

It has been generally accepted that metazoan genome replication proceeds by near synchronous firing of clusters of replication origins located adjacent to each other on the linear DNA sequence. However, the precise manner in which DNA replication proceeds within Mb-sized replication timing (RT) domains is not well understood.

In this manuscript, the authors analyzed the earliest RT domains on the human genome by NGS and performed wavelet analysis followed by a series of in-depth data analyses to understand how DNA replication proceeds within these domains. Specifically, the authors synchronized human U2OS cells by thymidine block&release followed by an L-mimosine block&release to label the earliest replicating sequences by a 30-min EdU pulse. This was followed by biotinylation, pull-down by streptavidin beads, and NGS to identify the sites of EdU incorporation on the U2OS genome. A series of experiments were performed to analyze 10–40, 40–70, 70–100, 100–130, and 0–130 min time points after release from the L-mimosine block.

The authors made the following findings:

1. The most prominent component of the early replicating signal was at the size of ~500 kb, which they defined as the Replicon Cluster Domains (RCDs).

2. RCDs were separated by ~1.2 Mb on average, and the number of RCDs was proportional to the size of the RT domains.

3. The >20-fold variation in replication peak height among RCDs was explained by the variation in the activation timing of RCDs rather than the difference in active origin density of RCDs.

4. RCDs gradually broadened in width over time at a pace consistent with the progression of two flanking forks at both ends of a given RCD proceeding at 1–1.5 kb/min.

5. The authors' data analysis suggested that most RCDs are still replicating internally even at the last time point, in parallel with the broadening of RCDs at both flanks by outward-moving forks.

6. They also addressed whether multiple RCDs within large early RT domains completed the replication of these domains by simple fusion of outward-moving forks flanking RCDs or by new initiation events within the 'valley' regions between RCD peaks. The authors' analysis suggested the latter, with later activating RCDs present in the 'valley regions' between early activating RCD peaks.

7. Finally, the authors tested the so-called 'domino' activation model of replication foci within single early RT domains. Their analyses suggested that adjacent RCDs tend to be activated much more frequently than by chance, which supported the 'domino' activation model.

It appeared to me that the analyses were done carefully and thoroughly to provide evidence supporting the points made by the authors. It is impressive to see how such simple NGS experiments can provide valuable insights into the regulation of DNA replication inside RT domains and underscores the importance of exploiting large-scale data sets through in-depth analyses. I am positive about the manuscript. I do have a few comments, though.

1. U2OS cells are aneuploid. Figure S2 G1 data indeed shows various copy number variations. I wonder if this could be a potential confounding factor in interpreting DNA replication data for some genomic regions.

2. A recent paper by the authors using optical mapping technology (Ref. 7) did not support the 'domino' activation model of replication initiation zones (IZ) and instead supported the stochastic activation model for neighboring IZs. It would help the readers if the authors discussed these differences and provided potential interpretations for the discrepancy.

3. It is interesting that while RCDs are gradually broadening in width at a pace consistent with the progression of two outward-moving forks at their flanks, most RCDs are still replicating internally even at the last time point of the experiment performed. Eventually, however, a time should come after which only the two outward-moving forks at the RCD flanks are synthesizing DNA in the absence of replication activities inside the RCD (as in Fig. 7a, the very bottom cartoon), i.e., timing transition region (TTR) replication phase. I wonder why the authors did not observe such a phase (as in Fig. 7a at the very bottom) in their experiment. For instance, is this due to variability in the timing of entry into the S-phase among cells after release from the L-mimosine block? Was 130 min not sufficient for the earliest replicating RCDs (replication foci) to complete their replication?

4. If the level of synchrony was sufficient, it seems possible to witness such a phase (as in Fig. 7a at the very bottom), especially for small early RT domains with only a single RCD.

5. (Optional) If the authors could see this, I am very curious as to the genomic locations of the two outward-moving forks at the RCD flanks in such phase based on the authors' analysis and how this is related to the TAD boundary positions because the authors previously reported the alignment of early borders of TTRs to TAD boundary positions. This would provide valuable insights into the 3D organization of TTRs and the relationship to replication activities. Hi-C data for U2OS cells are available (see links below), so this analysis seems feasible without additional experiments. Also, this might allow the authors to relate RCDs to

intra-TAD structures.
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM6552772
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4194463
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4194464

Minor points:

- Page 9, Ref.48 seems incorrect > a different Wang et al., paper (Mol Cell, 2021) and the same as Ref. 7, I believe.

- Figure 7b, I did not understand the figure very well. The vertical axis probably represents replication signal intensity (which should be described in the figure), but how could the height of the blue-green dashed line be identical to that of the blue line?

- Figure 8, I would like to see population Repli-seq data (as in Fig. 2b) for comparison to see that the valleys between peaks are part of the same early replicating RT domain.

- Page 14 & Figure 5a, how did the authors define the distance between adjacent RCDs? My prediction is end-to-end distance based on how things are described on page 14, but it could be the peak-to-peak distance between two adjacent RCDs. This should be described.

- Page 18, Although there was substantial a variability in > Although there was substantial variability in (typo)

- Page 19, line 1: this is at least in part part caused by > this is at least in part caused by

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have

significant reservations, as outlined above.

Author Response 14 Aug 2023

J Julian Blow

Main Point 1: "U2OS cells are aneuploid. Figure S2 G1 data indeed shows various copy number variations. I wonder if this could be a potential confounding factor in interpreting DNA replication data for some genomic regions."

This is indeed a possibility but is unlikely to have a significant effect on our interpretations. It would only affect the wavelet analyses in the rare cases where a translocation in the U2OS genome occurs within one of the 500 kb wavelet peaks. We have added the following text to page 12: "*It should be noted that our sequences have been mapped onto a normal human genome, but since U2OS have a number of chromosome rearrangements and copy number variations this could be a potential confounding factor in interpreting DNA replication data for a few genomic regions.*"

Main Point 2: "*A recent paper by the authors using optical mapping technology (Ref. 7) did not support the 'domino' activation model of replication initiation zones (IZ) and instead supported the stochastic activation model for neighboring IZs. It would help the readers if the authors discussed these differences and provided potential interpretations for the discrepancy.*"

The term "domino model" has been invoked to mean at least two different things. In the way we have used the term, it simply refers to the descriptive observation that nearby structural domains observed cytologically tend to fire sequentially - in other words, replication foci labeled with pulse - 60 minute chase (to complete the first foci) - second pulse - often are genetically close to each other. The prior optical replication mapping (ORM) paper (Wang et al, 2021) to which the reviewer refers could not address this version of the domino model because ORM is a single pulse label that cannot measure sequential events.

Moreover, the fibres in ORM average ~350kb, so they could not contain two adjacent domains. We acknowledge that there is a sentence in the ORM paper with the term "domino model" in it, but this was referring to a more strict version of the domino model in which a mechanism is invoked that forks emanating from one domain invade the adjacent domain activating initiation. This would create adjacent IZs of different sizes on the same single fibres (fork encroaching and stimulating initiation nearby) and we did not observe that. In fact, this mechanism was ruled out by Dimitrova and Gilbert, NCB 2000, who showed that by disabling the S phase checkpoint, one could get sequential foci firing in the proper order while completely inhibiting fork elongation.

To address this confusion in the paper, we have added to the sequential activation sub-section of the results the following clarification: "*The mechanism by which this occurs is not known, however, it is unlikely to occur by the forks emanating from one domain stimulating initiation of the adjacent domain because: a) disabling the S phase checkpoint in the absence of processive elongation of replication forks still results in the sequential firing of replication foci in the proper order (Dimitrova and Gilbert, NCB, 2000) and; b) optical replication mapping studies*

failed to find evidence for new bursts of DNA synthesis near longer tracks (Klein et. al., Science, 2021)."

Main Points 3 and 4: "*It is interesting that while RCDs are gradually broadening in width at a pace consistent with the progression of two outward-moving forks at their flanks, most RCDs are still replicating internally even at the last time point of the experiment performed. Eventually, however, a time should come after which only the two outward-moving forks at the RCD flanks are synthesizing DNA in the absence of replication activities inside the RCD (as in Fig. 7a, the very bottom cartoon), i.e., timing transition region (TTR) replication phase. I wonder why the authors did not observe such a phase (as in Fig. 7a at the very bottom) in their experiment. For instance, is this due to variability in the timing of entry into the S-phase among cells after release from the L-mimosine block? Was 130 min not sufficient for the earliest replicating RCDs (replication foci) to complete their replication? If the level of synchrony was sufficient, it seems possible to witness such a phase (as in Fig. 7a at the very bottom), especially for small early RT domains with only a single RCD*."

We did look for TTRs as suggested and also considered small early RT domains. As the reviewer suggests, this was confounded by the fact that the synchrony decays over the period of the analysis. Although our data would be consistent with TTRs, we decided not to try to analyse these as the data were too ambiguous.

Main Point 5. "*If the authors could see this, I am very curious as to the genomic locations of the two outward-moving forks at the RCD flanks in such phase based on the authors' analysis and how this is related to the TAD boundary positions because the authors previously reported the alignment of early borders of TTRs to TAD boundary positions. This would provide valuable insights into the 3D organization of TTRs and the relationship to replication activities. Hi-C data for U2OS cells are available (see links below), so this analysis seems feasible without additional experiments. Also, this might allow the authors to relate RCDs to intra-TAD structures.*"

We examined whether the 'isolated' RCDs analysed in Figures 6 and 7, where the RCD boundaries are fairly well defined, might correspond to published TAD boundaries as suggested. Despite some peaks reaching high association scores with overlapping TADs, none of these scores was found to be statistically significant at a 0.05 false discovery rate (FDR) level.

Minor points:

- *Page 9, Ref.48 seems incorrect > a different Wang et al., paper (Mol Cell, 2021) and the same as Ref. 7, I believe.*

  We've now used the correct reference.

- *Figure 7b, I did not understand the figure very well. The vertical axis probably represents replication signal intensity (which should be described in the figure), but how could the height of the blue-green dashed line be identical to that of the blue line?*

  We have clarified the description of the cartoon in the figure legend. We also appreciate the point that the blue-green dashed line should be higher than the blue

or green lines, so have modified the figure accordingly.

- *Figure 8, I would like to see population Repli-seq data (as in Fig. 2b) for comparison to see that the valleys between peaks are part of the same early replicating RT domain.*

  We have added the Repli-Seq (timing domain) data to the figure as requested. We appreciate this suggestion as one of the selections (chromosome 3, 111-121 MbP) spanned two timing domains, so we have replaced this with a new selection (chromosome 7 39-49 Mbp).

- *Page 14 & Figure 5a, how did the authors define the distance between adjacent RCDs? My prediction is end-to-end distance based on how things are described on page 14, but it could be the peak-to-peak distance between two adjacent RCDs. This should be described.*

  The distance was between the peaks as defined by the 500 kb wavelet analysis; the text has been amended to clarify this ("Figure 5a shows that the peak separation…").

- Page 18, Although there was substantial a variability in > Although there was substantial variability in (typo)

  Corrected.

- Page 19, line 1: this is at least in part part caused by > this is at least in part caused by

  Corrected.

Competing Interests: No competing interests were disclosed.