Supplementary information

Replication stress induces amplification of early replicating loci in the human genome

Fang Ji[1,2,3]#, Hongwei Liao[1,2]#, Haoyu Tang[1,2]#, Liujian Ouyang[2], Menglin Qiu[1,2], Yan Chen[1,2], Yingfei Huang[2], Jixuan Liu[1,2], Wen Li[2], Zhihua Chen[2], Ian David Hickson[4]* and Songmin Ying[1,2,3]*

[1]Department of Pharmacy, Center for Regeneration and Aging Medicine, the Fourth Affiliated Hospital of School of Medicine, and International School of Medicine, International Institutes of Medicine, Zhejiang University, Zhejiang-Denmark Joint Laboratory of Regeneration and Aging Medicine; Yiwu, 322000, China.

[2]Key Laboratory of Respiratory Disease of Zhejiang Province, Department of Respiratory and Critical Care Medicine, Second Affiliated Hospital of Zhejiang University School of Medicine; Hangzhou, Zhejiang, 310009, China.

[3]Department of Pharmacology, Zhejiang University School of Medicine; Hangzhou, Zhejiang, 310009, China.

[4]Center for Chromosome Stability, Department of Cellular and Molecular Medicine, University of Copenhagen; 2200 Copenhagen N, Denmark.

# These authors contributed equally to this work.

*Corresponding author. Email: yings@zju.edu.cn (S.Y.) or iandh@sund.ku.dk (I.D.H.)

Materials and Methods

Cell culture and manipulations

U2OS and HeLa cells were grown in Dulbecco's modified Eagle's medium containing 10% fetal bovine serum (FBS). Human bronchial epithelial (HBE) cells were cultured in RPMI 1640 medium containing 10% FBS. Mouse embryonic fibroblasts (MEF) were maintained in Dulbecco's modified Eagle's medium with 10% FBS, 1% Non-Essential Amino Acids (Sigma) and 1% GlutaMAX (Sigma). All media were supplemented with 1% penicillin-streptomycin (Invitrogen). Cells were maintained at 37˚C in an atmosphere of 5% CO_2.

For arresting DNA replication, cells were grown to 60-80% confluence and treated with 3 mM hydroxyurea (Sigma, H8627) or 4 μM aphidicolin for the times indicated. Cells were subsequently washed three times with PBS, and released into pre-warmed medium containing 20 μM EdU for a further 30 min. As for the control group, cells were only labelled with 20 μM EdU for 30 min. Finally, cells were fixed with 90% ice-cold methanol for at least 20 min and subjected to EdUseq.

Mapping of early origins

U2OS and HeLa cells were treated with nocodazole for 12 hrs to induce mitotic arrest. Mitotic cells were then manually collected by shake-off and re-seeded in warm medium containing 20 μM EdU and 4 mM hydroxyurea for 16 hrs to label nascent DNA in early S cells. Afterwards, cells were harvested and subjected to EdUseq.

Repli-Seq assay

Asynchronous U2OS cells were incubated with media containing 20 μM EdU for 0.5 hr, before being harvested and fixed with 90% ice-cold methanol for 20 min. After permeabilization, cells were stained with DAPI at room temperature for 1 hr. Cells were then sorted into different fractions according to their DNA content. Finally, sorted cells were subjected to EdUseq.

EdUseq

Cells were permeabilized with PBS containing 0.5% Triton X-100 for 20 min. And incorporated EdU was attached to a biotin linker (biotin-azide) (Thermo Fisher, B10184) by click reaction. Then, cells were lysed and subjected to DNA extraction. DNA fragmentation was achieved using a Bioruptor (Diagenode). Then, biotinylated DNA was enriched through biotin-streptavidin reaction using Dynabeads MyOne Streptavidin C1 Beads (Thermo Fisher Scientific, 65001) and amplificated with KAPA Hyper Prep Kit (KK8502) following the manufacturer's instructions. Finally, constructed DNA library was sequenced at AnnoroadGene Technology Co., Ltd. (Beijing, China) using the NovaSeq X PLUS.

Fluorescence-activated cell sorting

Cells were fixed by 90% ice-cold methanol and stained with DAPI at room temperature for 1 hr. Stained cells were sorted into different fractions according to their DNA content using a BD FACS Aria II SORP cell sorter.

Cleavage Under Target & Tagmentation (CUT&TAG)

CUT&TAG was performed using a NovoNGS CUT&TAG 2.0 High-Sensitivity Kit (Novoprotein, N259-YH01) following the manufacturer's instructions. Briefly, up to 100,000 live cells were incubated with an appropriate volume of ConA beads for 10 minutes. The bead-cell complexes were then sequentially incubated with the primary and secondary antibodies for 2 hours at room temperature on a rotator. After three washes with pre-cooled secondary antibody buffer, the bead-cell complexes were incubated with 1 μL ChiTag 2.0 transposome in 100 μL ChiTag buffer for 1 hour at room temperature on a rotator. Tagmentation was carried out by incubating the bead-cell complexes with 300 μL tagmentation buffer at 37°C for 1 hour on a rotator. To recover DNA, the bead-cell complexes were washed twice with PBS, and then lysed in 50 mM Tris (pH 8.0), 1% SDS, and proteinase K at 50°C overnight. DNA was then extracted using the phenol/chloroform method and eluted in 40 μL of pre-warmed elution buffer. The eluted DNA was used for library preparation and purified with AMPure XP (Beckman Coulter, A63880). Finally, the purified DNA library was sequenced at AnnoroadGene Technology Co., Ltd. (Beijing, China) using the NovaSeq X PLUS (Illumina platform).

Primary antibodies used in this study were as follows: mouse anti-phospho-Histone H2A.X (Ser139) antibody (Millipore, 05-636-l), rabbit anti-RAD51 antibody (abcam, ab133534,1:1000), rabbit anti-Histone (acetyl K27) antibody (abcam, ab4729,1:1000), rabbit anti-ORC1 antibody (abcam, ab85830,1:1000). Secondary antibodies used in this study were as follows: Goat pAb to Rabbit (abcam, ab6702), Goat pAb to Mouse (abcam, ab6708).

G1 cells collection for ORC1 CUT&TAG

Asynchronous U2OS cells were incubated with media containing 7 μM RO3306 for 18 hrs. Then, cells were released into M phase by washing three times with prewarmed PBS and incubated for 30 min. Afterwards, mitotic cells were manually shaken off and re-seeded onto poly-lysine-coated glass slides (Sigma-Aldrich). After an additional 5-hour incubation, G1 cells were harvested and fixed for ORC1 CUT&TAG.

RNA interference

Transfections were carried out with the Lipofectamine™ RNAiMAX transfection reagent (Thermo Fisher Scientific, Cat. # 13778150), according to the manufacturer's instructions. Transfections were performed using 25 nM siRNAs, and experiments were performed approximately 60 hrs post-transfection. SiRNAs were designed as followed: PIF1: (5'-GCAGGAACAGGGAAGUCAUTT-3'); SLX4: (5'-GACCAGAAGCAGAGAAUAATT-3'); negative control siRNA: (5'-UUCUCCGAACGUGUCACGUTT-3'). The knockdown efficiency was monitored by QPCR and western blot analysis.

QPCR and Western blotting

As for QPCR, cells were collected after transfection with siRNA for 60 hrs and were then immediately lysed using Trizol (Invitrogen). Total mRNA was extracted and reverse-transcribed to cDNA following the manufacturer's instructions. Transcripts of interest were quantified with real-time qPCR amplification. The primers used were as followed: PIF1-F: (5'-CATCCACAAGAGCCAAGGCAT-3'); PIF1-R: (5'-GGTGGCATAGAAGTGCAGCA-3'); SLX4-F: (5'-TTGGTCCTACAGCGAATGCAG-3'); SLX4-R: (5'-CATGTGCCGATGCTCCTACC-3'); GAPDH-F: (5'- CTGGGCTACACTGAGCACC-3'); GAPDH-R: (5'-AAGTGGTCGTTGAGGGCAATG-3').

As for western blotting, whole cells were lysed in RIPA buffer complemented with protease inhibitor (Cat #11873580001, Roche) and PMSF (Phenyl-Methanesulfonyl Fluoride Solution, Sigma Aldrich). After sonication and protein quantification, samples were heated at 95 ˚C for 5 min in SDS loading buffer (1:1 ratio). An aliquot of sample was run on the SDS–PAGE gel and then transferred to PVDF membranes. After being blocked with 5% milk in TBST buffer for 1 hr at room temperature, membranes were incubated overnight with appropriate dilutions of primary antibody in 5% milk at 4 °C. Following three washes using TBST (5 min each), the membranes were incubated with appropriate dilutions of secondary antibody in 5% milk for 1 hr at room temperature and washed for three times using TBST. Detection was achieved by Odyssey CLx Imager and results were analyzed using FIJI. Primary antibodies used were: mouse anti-SLX4 antibody (Abcam, ab169114, 1:1000), mouse anti-PIF1 antibody (Santa Cruz, #sc-48377, 1:1000), rabbit anti-Actin antibody (ABclonal, AC026). Secondary antibodies used were: Goat anti-rabbit IgG (H+L) Dylight 680 (EarthOx, E032720), Goat anti-mouse IgG (H+L) Dylight 680 (EarthOx, E032710).

Genome alignment and peak calling

Sequencing reads were trimmed and aligned to hg38 human reference genome or mm10 mouse reference genome using Trim Galore and Bowtie2, respectively. PCR duplicates were removed by SAMtools and clean reads were transformed to BigWig format using deepTools2. And EdUseq peaks were called using MACS2 with default settings.

Definition of replication restart hotspots (RHs) and control sites (CSs)

For each 50 kb non-overlapping windows, read coverage was first computed using reads per kilobase (RPK). And the EdUseq signal fold change of each 50 kb non-overlapping windows was calculated by dividing the read coverage of experiment group to that of control group. The threshold of fold change for defining replication restart hotspots (RHs) was first screened with a differential function for fold change distribution that identified potential background cutoffs, and then refined based on fitness to EdU peaks inspected in IGV. Eventually, 1658 regions were defined as RHs at the threshold of 7. And the same amount of random counterpart (named control sites (CSs)), with Y chromosome, telomere and centromere sequence excluded, were generated for comparison. As for HBE cells, 3516 regions whose fold change is more than 32 were defined as RHs. As for MEF cells, 1940 regions whose fold change is more than 2.5 were defined as RHs.

Replication timing analysis

Sequencing reads from Repli-Seq assay were computed as described above. For IGV visualization, log2 ratios of the early vs late read counts were calculated for each nonoverlapping 5-kb window as previously reported[14]. For classification of early, middle and late-replicating regions, the replication time estimator S50, defined as the fraction of the S phase ($0 < S50 < 1$) at which 50% of DNA was replicated in a defined genome region, was calculated as previously described[15].

Definition of re-replication sites (reRSs) and control sites (CSs)

First, the CPM (counts per million) was computed for each 50 kb non-overlapping window for control and treated late S/G2/M samples. Then, windows whose CPM is in the top 50% in treated late S/G2/M sample and in the bottom 50% in control late S/G2/M sample were selected for further analysis. Next, the EdUseq signal fold change of selected windows was calculated by dividing the CPM of treated group to that of control group. The threshold for reRSs calling was first screened with a differential function for fold change distribution that identified potential background cutoffs, and then refined to be 4 based on fitness to EdU signals inspected in IGV. For comparison, we generated random counterpart (named control sites (CSs)) across the genome, with Y chromosome, telomere and centromere sequence excluded. The size and number of CSs in each chromosome were paired to reRSs. The fold change threshold for defining APH-induced reRSs in U2OS cells was 5, and 3 for HU-induced reRSs in HeLa cells, using the same criterion described.

Genomic annotations and gene ontology analysis

R package ChIPseeker was used to annotate EdUseq peaks to genetic elements and the closest genes. And gene ontology analysis was performed with R package clusterProfiler.

Calculation of GC content, AT content, CpG number and motif identification

To quantify GC and AT content, bedtools was first used to extract DNA sequences and the percentage of overall G and C contents or overall A and T contents for each RH and CS was then calculated. To compare the CpG coverage between RHs and CSs, CpG information was first applied from the following website: https://github.com/Bioconductor/BiocWorkshops/blob/master/100_Morgan_RBiocForAll/CpGislands.Hsapiens.hg38.UCSC.bed. And the length covered by CpG for each RH and CS was then calculated. DNA sequence motifs were identified by findMotifsGenome.pl of homer.

Pearson correlation analysis

Peaks were independently called for each replicate using MACS2 with default parameters, and a union peak set was generated by merging all peak regions detected in either replicate using bedtools. For each peak in the union set, EdU–seq signal intensities from both replicates were quantified and normalized as counts per million (CPM). The resulting signal matrix was then used to compute Pearson correlation coefficients. Higher Pearson correlation coefficients indicate stronger reproducibility between replicates.

Analysis of ATAC-seq, H3K27ac ChIP-seq, CUT&TAG data and deposited data

ATAC-seq and H3K27ac ChIP-seq was performed according to the manufacturer's protocol. And obtained DNA library was sequenced at Jinan Ruizhi Biotechnology Co., Ltd. (Shandong, China) using the HiSeq Xten (Illumina platform). Sequencing reads from ATAC-seq, H3K27ac ChIP-seq, CUT&TAG, and ChIP-seq of downloaded histone modifications were subjected to trimming, aligning and duplicate removal using Trim Galore, Bowtie2 and SAMtools, respectively. DeepTools2 was used to transform bam files into BigWig format. The counts per million (CPM) algorithm was used for genome-wide analysis of related signals in each reRS or CS.

RNA-seq data analysis

RNA from asynchronously growing U2OS cells was sequenced commercially by AnnoroadGene Technology Co., Ltd. (Beijing, China). Reads were trimmed using Trim Galore with default settings and mapped to the human genome (hg38) by HISAT2, and the reads were counted using FeatureCounts for exons. We first attained the list of all human genes including their Entrez ID and location information from UCSC.hg38.known genes (version 3.10.0). And then calculated the overlapped length of each gene adjacent to each reRS or CS. Overlapped genes were defined as if they meet any of the following criteria: 1) overlapped length was more than 50% of the width of sites or the width of the gene itself; 2) the overlapping region was more than 250 kb; 3) the gene contains at least two sites. The reads per kilobase per mega reads (RPKM) were calculated to represent the transcription signal for CSs-overlapped genes (CSs-OV) and reRSs-overlapped genes (reRSs-OV).

Analysis of copy number variations

Whole genome sequencing (WGS) data from HU-treated U2OS cells were produced by AnnoroadGene Technology Co., Ltd, Beijing, China. Reads were first trimmed by Trim Galore before being aligned by BWA. Control-FREEC was used to call estimated copy number with default setup file. Loss CNV was defined by estimated copy numbers less than 1 and gain CNV was defined by estimated copy numbers more than 3. The number of gain or loss CNVs contained by reRSs or CSs was calculated.

Osteosarcoma-specific somatic CNV loci were obtained from the Catalogue of Somatic Mutations in Cancer (COSMIC), and pan-cancer CNV profiles were collected from the Cancer Genome Atlas (TCGA), covering six cancer

types: STAD (Stomach Adenocarcinoma); LUAD (Lung Adenocarcinoma); LIHC (Liver Hepatocellular Carcinoma); ESCA (Esophageal Carcinoma); COAD (Colon Adenocarcinoma); and CESC (Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma). The genomic distance between each CNV and the nearest reRS was calculated using bedtools closest (v2.28.0). CNVs were then classified into three categories based on their distances: $\leqslant$50 kb, 50-500 kb, and >500 kb. The proportion of CNVs in each category was calculated. For comparison, 1000 sets of random genomic loci, matched in both number and size to the observed CNVs, were generated and analyzed in the same manner.

Analysis of constrained elements

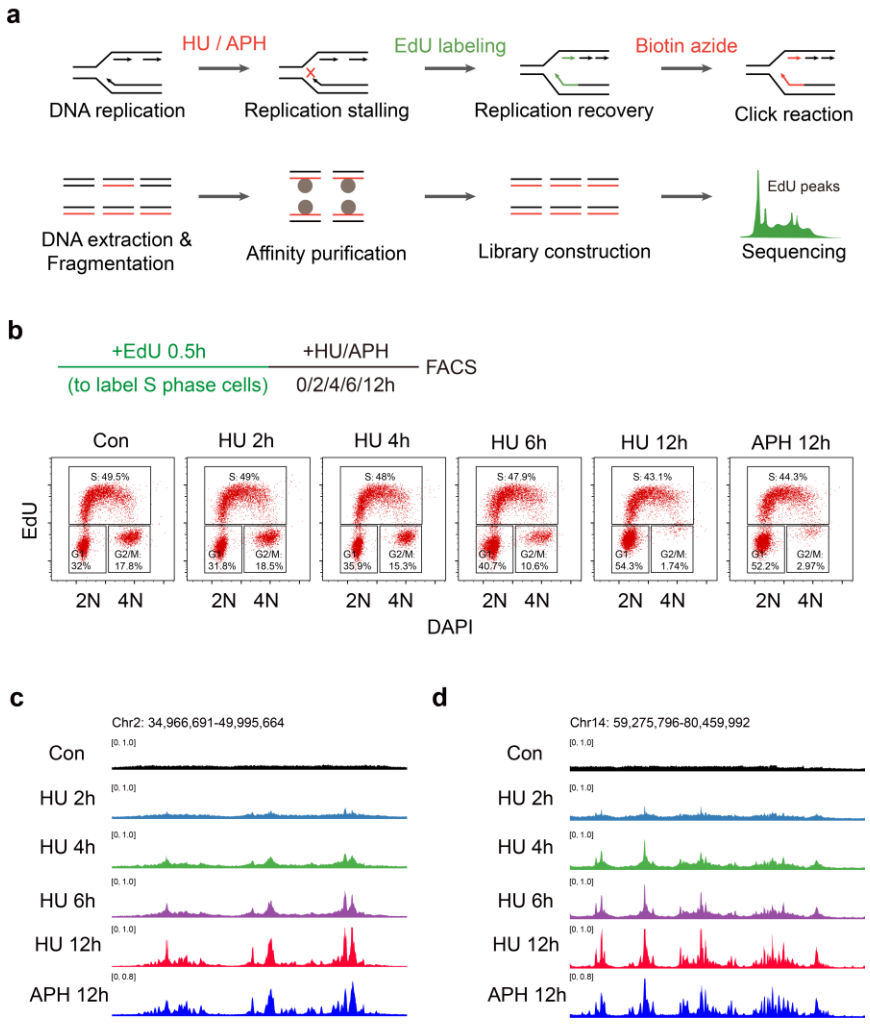
Constrained elements (CEs) were identified by genomic evolutionary rate profiling (GERP) as previously described[10]. The number of CEs contained by reRSs or CSs was first calculated. The sum of CE score of each contained CE was used to represent the CE level for each reRS or CS.

Statistical analysis

Statistical analysis was performed with Wilcoxon rank-sum test. Pearson correlation coefficient was used to measure the linear relationship between two variables.

Data visualization

pyGenometracks was used to create all of the representative genomic profiles. The aggregation plots were created using deepTools2, and the Circos plots were created using Circos to display the entire genome. Ggplot2 was used to produce the scatter plots and violin plots. DOSE was used to generate the dotplot for gene ontology analysis. The Venn diagram was drawn using VennDiagram.

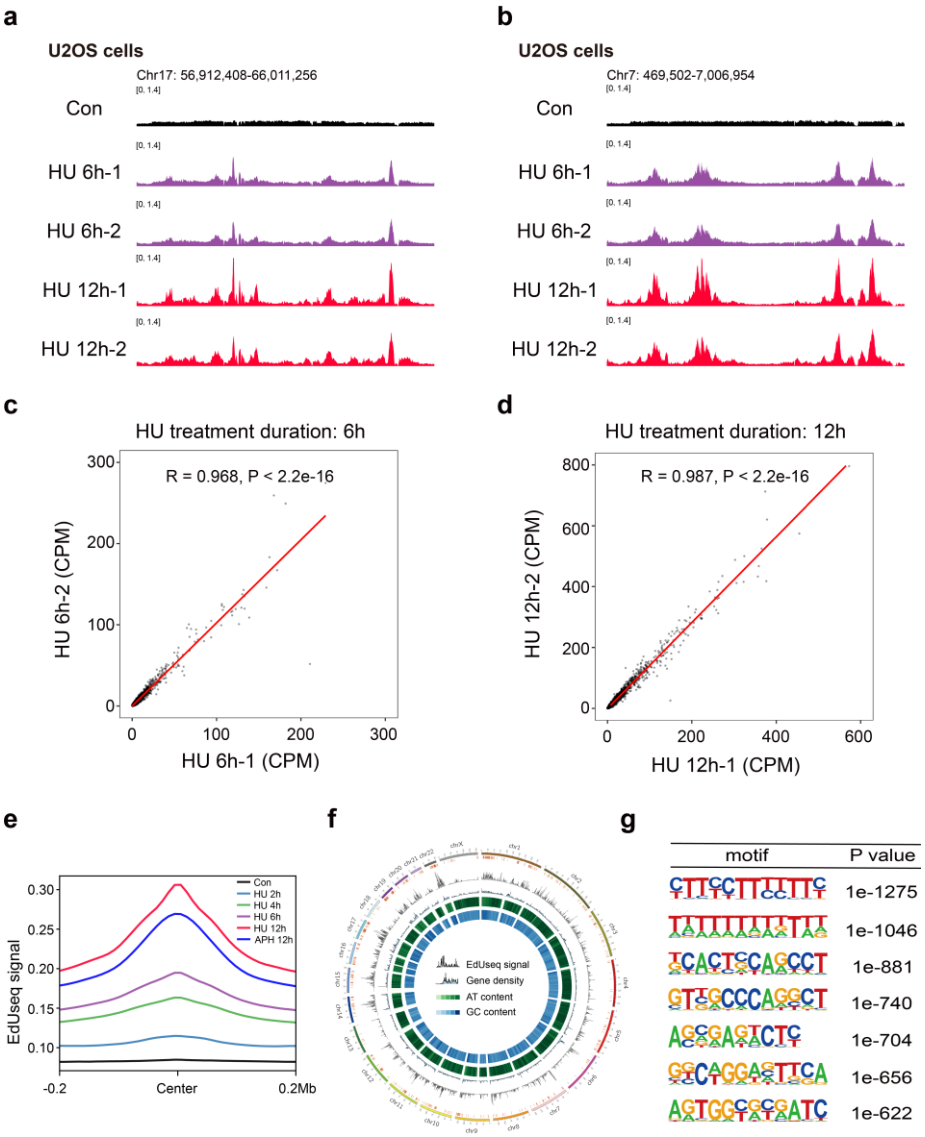


Supplementary Fig. S1.

a Schematic diagram for genome-wide mapping of replication restart using EdUseq.

b U2OS cells treated with HU or APH for different durations were analyzed by flow cytometry for DNA content and EdU staining.

c-d Representative views of genomic distribution of EdUseq signals of U2OS cells in chromosome 2 and 14.



Supplementary Fig. S2.

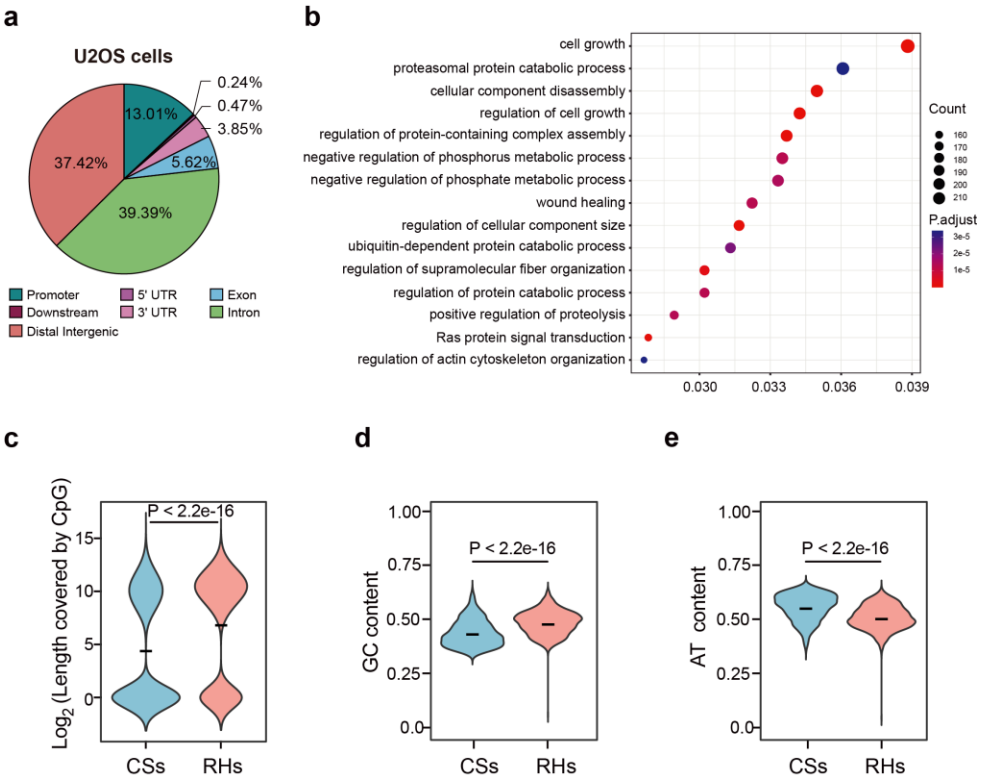
a-b Genomic view of EdUseq signals showing high reproducibility between replicates.

c-d Pearson correlation analysis for the two replicates indicated.

e An aggregate plot showing EdUseq signals around the centers of peaks.

f Circos plot of the whole-genome distribution of EdUseq signals, gene density, and AT/GC contents. Darker green and blue indicate higher AT content and GC content, respectively.

g DNA sequence motifs identified in EdUseq peaks. P value indicated statistical significance.

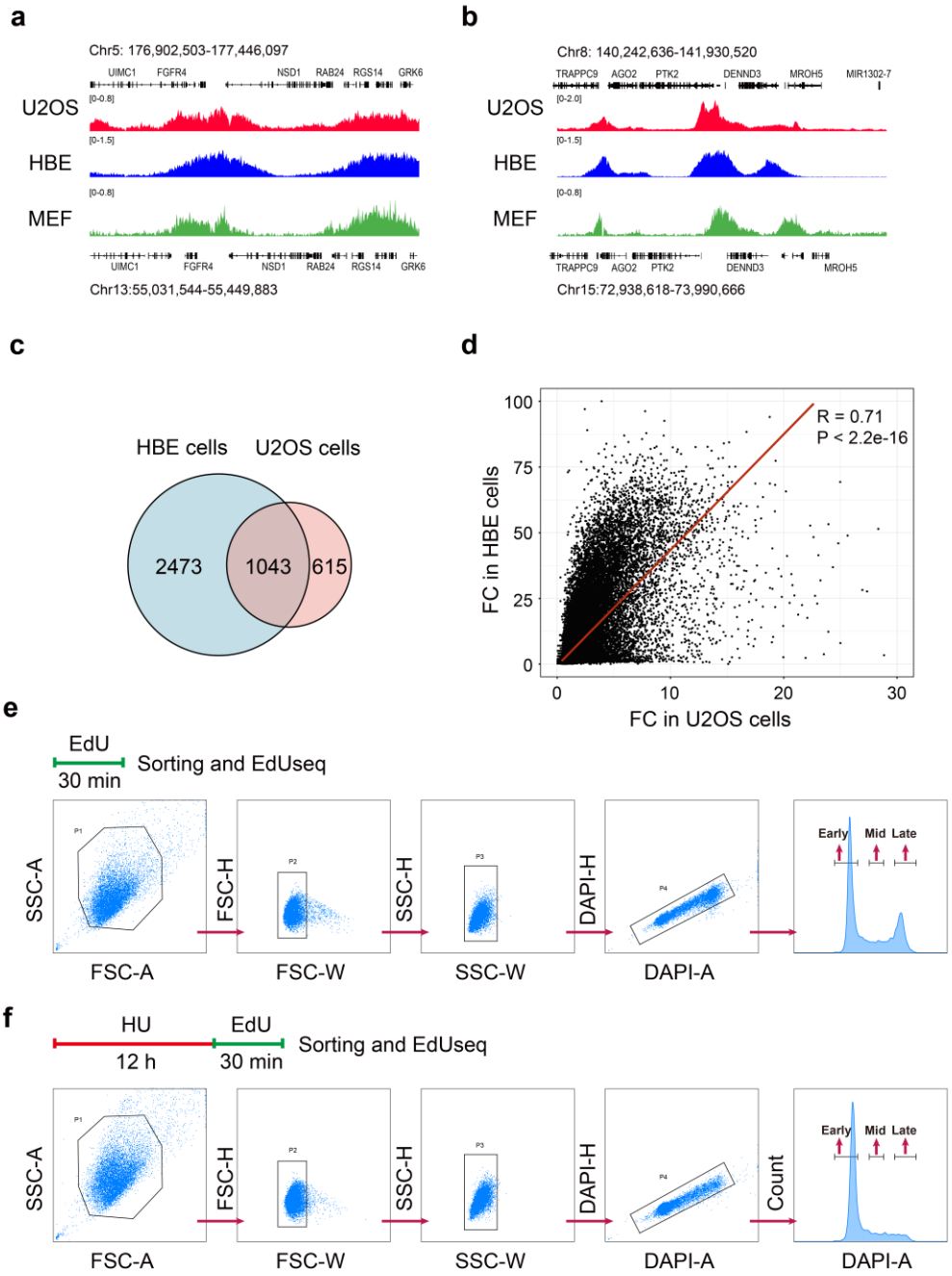


Supplementary Fig. S3.

a Distribution of EdUseq peaks across the genetic elements.

b Top 15 GO terms for EdUseq peak-related genes.

c-e The CpG level, overall GC content and AT content in CSs and RHs. P value was determined using the Wilcoxon rank-sum test. The solid line indicates the median value.



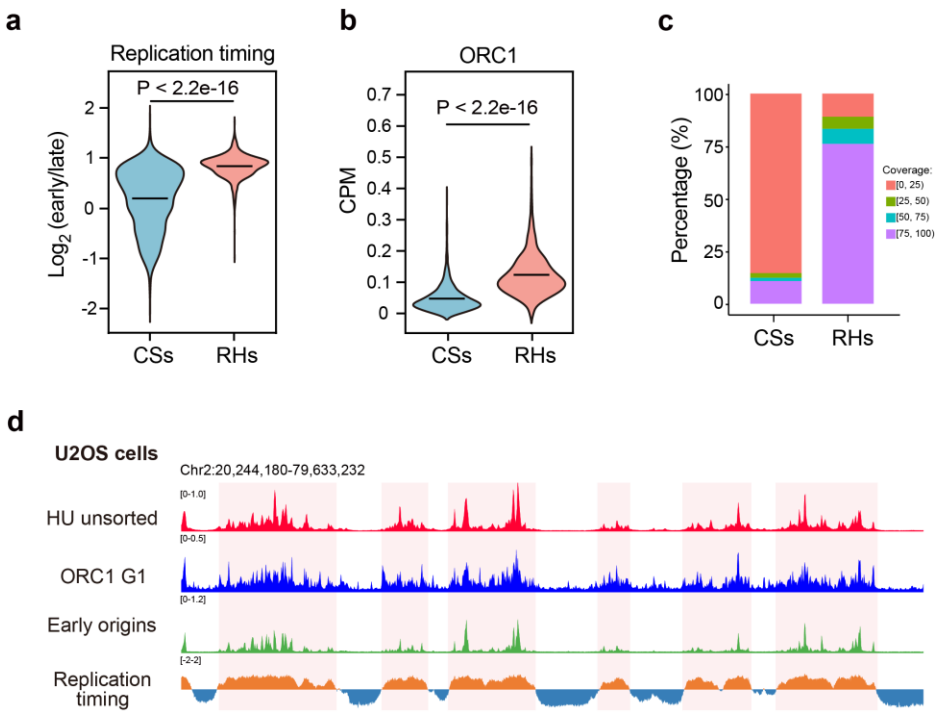
Supplementary Fig. S4.

a-b Representative genomic views of the EdUseq signal in human U2OS cells, HBE cells, and mouse embryonic fibroblasts.

c Venn diagram showing the overlap of RHs between HBE cells and U2OS cells.

d The EdUseq signal fold change (FC) of 50 kb non-overlapping regions across the genome in HBE and U2OS cells.

e-f, The sorting strategy used to separate different fractions of cells in the absence or presence of HU treatment, respectively.

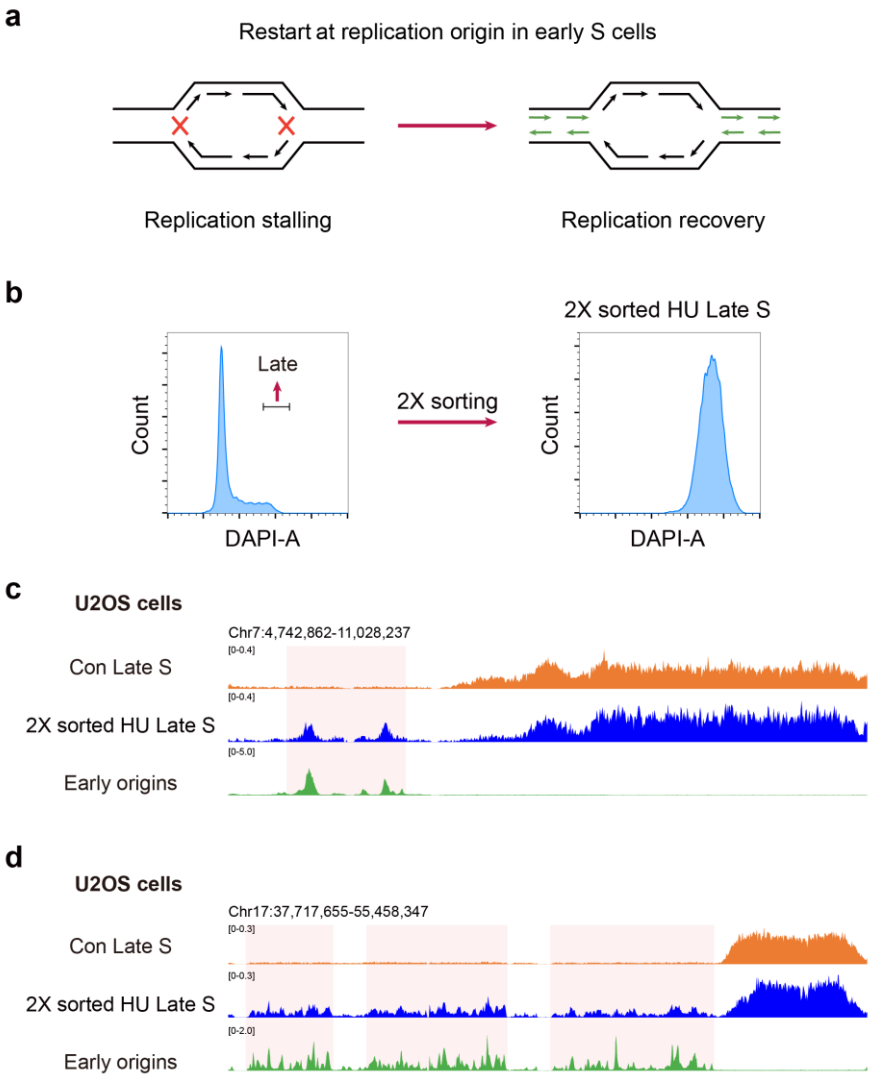


Supplementary Fig. S5.

a-b Quantitative comparison of replication timing and ORC1 signal between CSs and RHs. P values were determined by Wilcoxon rank-sum test. The solid line indicates the median value.

c Statistical result of the percentage of CSs or RHs overlapped with early origins. Coverage rate is calculated by the overlapped length divided by the length of CSs or RHs.

d A representative view of a 60-Mb genomic region with tracks showing EdUseq signals of U2OS cells released from 12h-treatment with HU, ORC1 signal at G1 phase, early origins, and replication timing. Yellow and blue peaks in the track of replication timing indicate early-replicating and late-replicating regions, respectively.

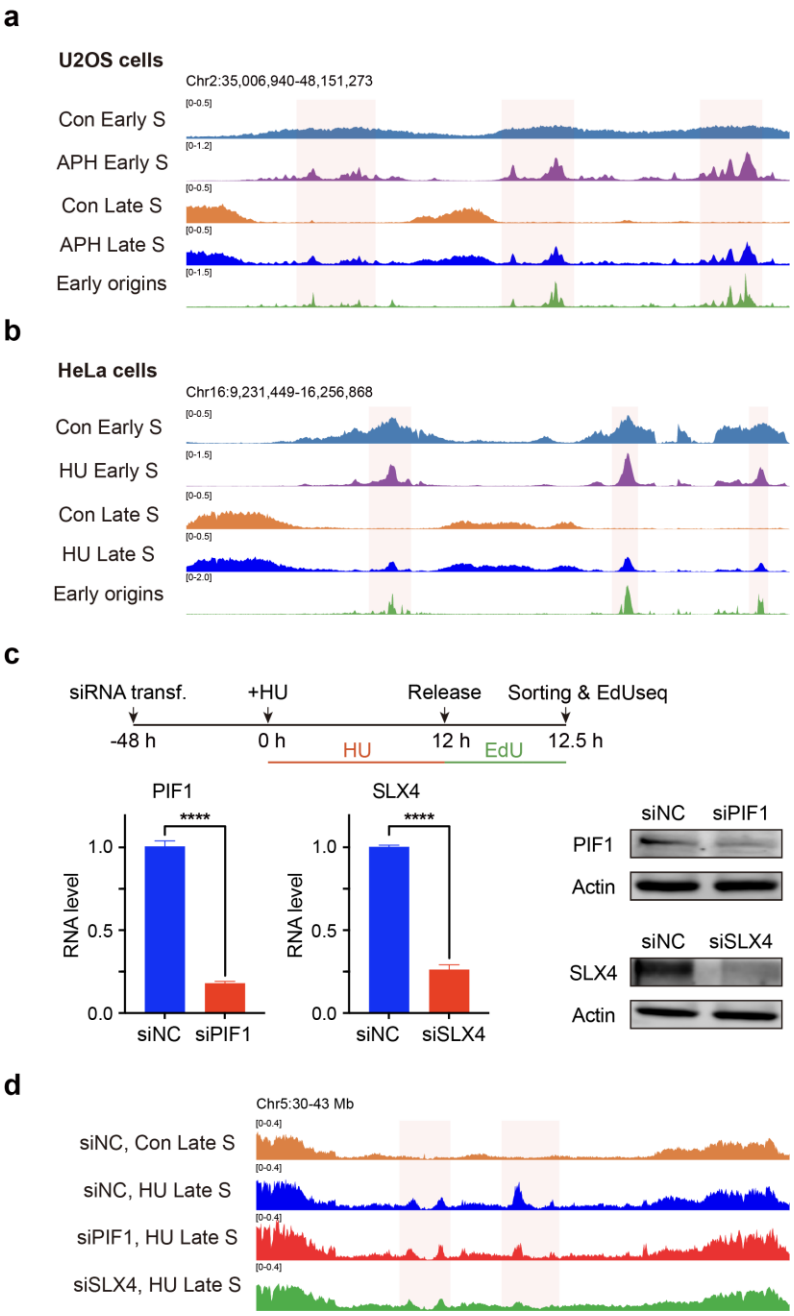


Supplementary Fig. S6.

a Schematic view of replication restart at origins in early S cells.

b To check the fractionation quality, the 2X sorted cells, already stained with DAPI, were directly re-analysed by flow cytometry.

c-d Representative genomic views of the EdUseq signals of sorted U2OS cells treated as indicated for different tracks.

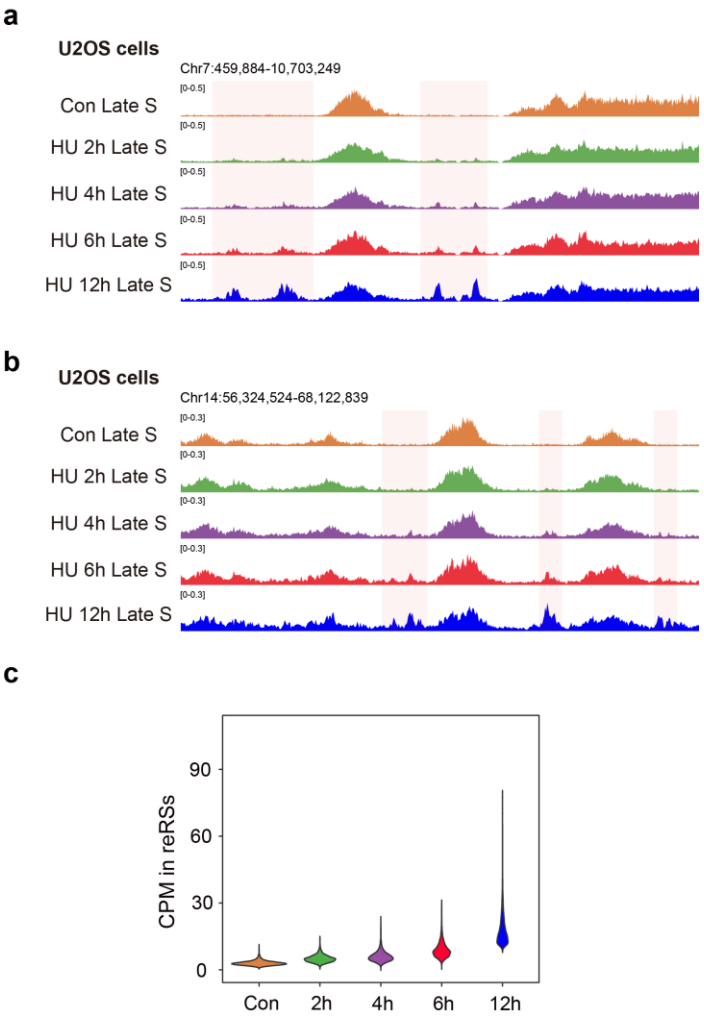


Supplementary Fig. S7.

a-b A representative genomic view of the EdUseq signals of sorted U2OS or HeLa cells treated as indicated for different tracks.

c Upper panel, experimental scheme for performing EdUseq. Lower panel, QPCR and immunoblot analysis for U2OS cells transfected with negative-control short interfering siRNA (siNC) or siRNA that targets PIF1 (siPIF1) or SLX4 (siSLX4).

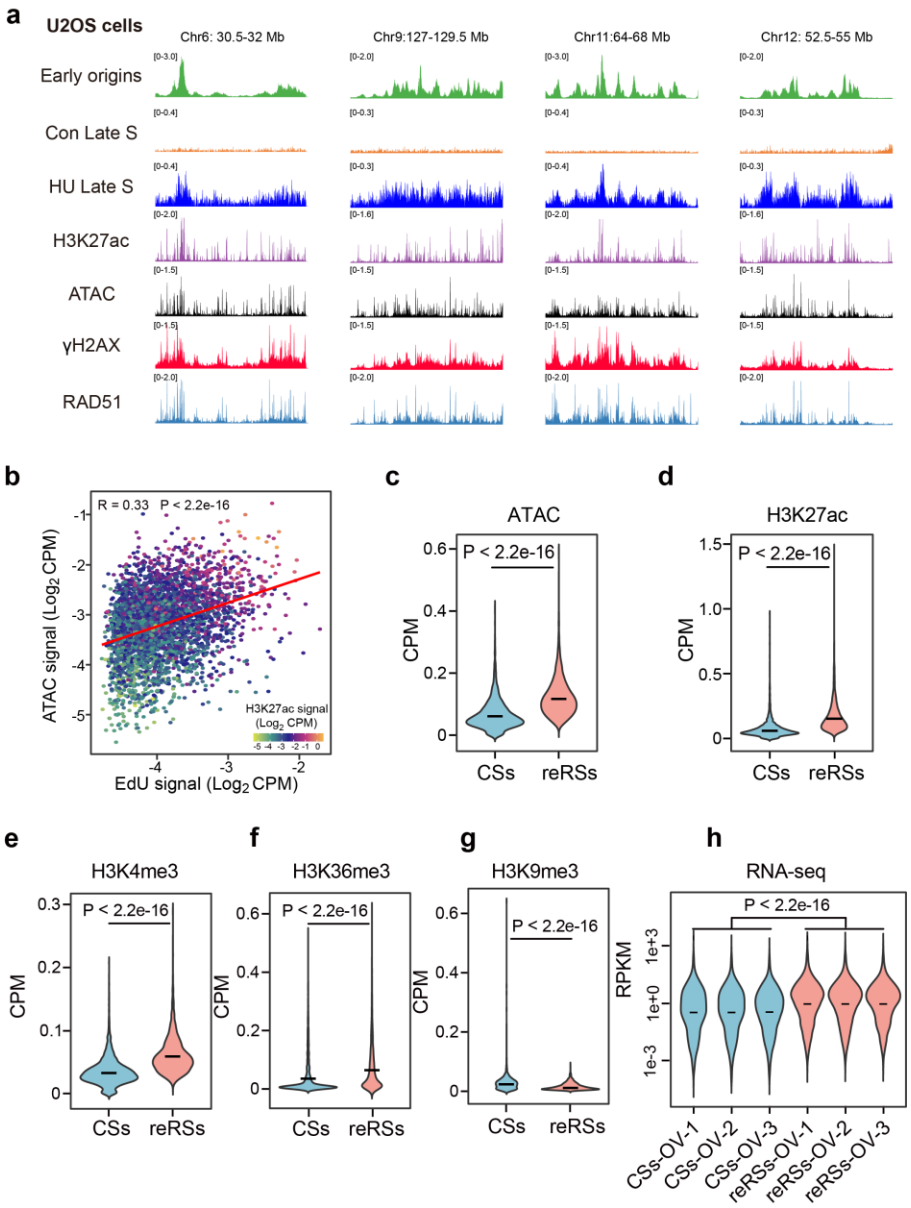
d Representative genomic view of EdUseq signals after depletion of the proteins indicated.



Supplementary Fig. S8.

a-b Genomic view of representative regions showing the EdUseq signals at reRSs following HU treatment for different durations.

c The quantification of EdUseq signals at reRSs.



Supplementary Fig. S9.

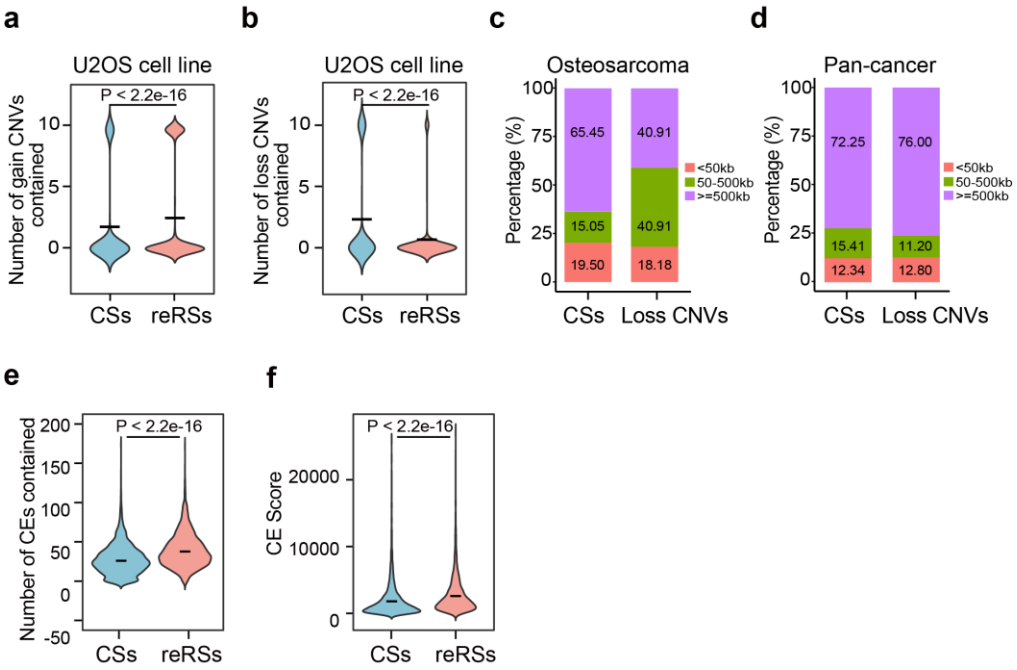
a Genomic view of four representative regions showing the colocalization of reRSs with early origins, hotspots of DNA damage response proteins, and epigenetic features.

b A scatter plot displaying the signals of EdUseq, ATAC-seq and H3K27ac ChIP-seq for defined reRSs.

c-d Violin plots showing the signals of ATAC-seq, H3K27ac in CSs versus reRSs. P value was determined using the Wilcoxon rank-sum test. The solid line indicates the median value.

e-g Violin plots comparing the levels of H3K4me3, H3K36me3 and H3K9me3 in CSs and reRSs, respectively. P value was determined using the Wilcoxon rank-sum test. The solid line indicates the median value.

h The comparison of transcription level between CSs-overlapped genes (CSs-OV) and reRSs-overlapped genes (reRSs-OV). Data are from three parallel experiments. P value was determined using the Wilcoxon rank-sum test. The solid line indicates the median value.



Supplementary Fig. S10.

a-b Violin plots showing the number of gain CNVs and loss CNVs in CSs versus reRSs. P values were determined by Wilcoxon rank-sum test. The solid line indicates the median value.

c The percentages of osteosarcoma-specific loss CNVs and their matched random sites (CSs), at different distances from the nearest reRS were calculated.

d The percentages of pan-cancer-associated loss CNVs, as well as their matched random sites (CSs), at different distances from the nearest reRS were calculated.

e Comparison of the number of constrained elements (CEs) contained by CSs versus reRSs. P values were determined by Wilcoxon rank-sum test. The solid line indicates the median value.

f CE score of CSs and reRSs. P values were determined by Wilcoxon rank-sum test. The solid line indicates the median value.

References

14. Marchal, C. et al. Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq. Nat Protoc 13, 819-839 (2018).

15. Chen, C. L. et al. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res 20, 447-457 (2010).