3D chromatin connectivity underlies replication origin efficiency in mouse embryonic stem cells

Karolina Jodkowska[1,10,11,†], Vera Pancaldi [3,9,†], Maria Rigau[3,7,8,‡], Ricardo Almeida[2,‡], José M. Fernández-Justel[2], Osvaldo Graña-Castro[4,6], Sara Rodríguez-Acebes[1], Miriam Rubio-Camarillo[4], Enrique Carrillo-de Santa Pau[5], David Pisano[4], Fátima Al-Shahrour[4], Alfonso Valencia[3], María Gómez [2,*] and Juan Méndez [1,*]

[1]DNA Replication Group, Molecular Oncology Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain, [2]Functional Organization of the Mammalian Genome Group, Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Madrid, Spain, [3]Computational Biology Life Sciences Group, Barcelona Supercomputing Center (BSC), Barcelona, Spain, [4]Bioinformatics Unit, Structural Biology Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain, [5]Computational Biology Group, IMDEA Food Institute, Madrid, Spain, [6]Institute of Applied Molecular Medicine (IMMA-Nemesio Díez), San Pablo-CEU University, Boadilla del Monte, Madrid, Spain, [7]MRC London Institute of Medical Sciences, Du Cane Road, London W12 0NN, UK, [8]Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, Hammersmith Hospital Campus, Du Cane Road, London W12 0NN, UK, [9]Université de Toulouse, Inserm, CNRS, Université Toulouse III-Paul Sabatier, Centre de Recherches en Cancérologie de Toulouse, Toulouse, France, [10]Laboratory of Functional and Structural Genomics, Center of New Technologies, University of Warsaw, Warsaw, Poland and [11]Centre for Advanced Materials and Technologies, Warsaw University of Technology, Poleczki 19, 02-822 Warsaw, Poland

ABSTRACT

In mammalian cells, chromosomal replication starts at thousands of origins at which replisomes are assembled. Replicative stress triggers additional initiation events from 'dormant' origins whose genomic distribution and regulation are not well understood. In this study, we have analyzed origin activity in mouse embryonic stem cells in the absence or presence of mild replicative stress induced by aphidicolin, a DNA polymerase inhibitor, or by deregulation of origin licensing factor CDC6. In both cases, we observe that the majority of stress-responsive origins are also active in a small fraction of the cell population in a normal S phase, and stress increases their frequency of activation. In a search for the molecular determinants of origin efficiency, we compared the genetic and epigenetic features of origins displaying different levels of activation, and integrated their genomic positions in three-dimensional chromatin interaction networks derived from high-depth Hi-C and promoter-capture Hi-C data. We report that origin efficiency is directly proportional to the proximity to transcriptional start sites and to the number of contacts established between origin-containing chromatin fragments, supporting the organization of origins in higher-level DNA replication factories.

INTRODUCTION

Replication origins are fundamental elements for genomic stability, and their precise number, position and regulation in different organisms has been subject to decades of investigation (1,2). Different approaches have been used to identify origins in mammalian cells at the genome-wide level, including the isolation of 'short nascent DNA strands' (3–9); capture and sequencing of origin-containing replication 'bubbles' (10); analysis of the strand distribution of Okazaki

---

fragments (11,12); sequencing of newly-synthesized labeled DNA (13–16); chromatin immunoprecipitation of origin-binding proteins (17–20) and more recently, fluorescent visualization of initiation sites in single DNA molecules (21). Some of these approaches identify discrete initiation sites (origins) while others detect broader initiation zones (IZs) containing multiple origins. This apparent discrepancy has led to frequent debates in the field, including whether origin activation is purely deterministic (i.e. all cells in a population activate a same set of biochemically predetermined origins) or include a stochastic component. In the latter case, each cell in the population may use a different subset of potential origins located within larger IZs, and the selection of one origin or another is regulated by rules of probability. Regardless of the precise position and mode of activation of initiation sites, genome duplication follows a replication timing (RT) program that correlates with large-scale chromatin folding and is needed to maintain the landscape of histone modifications through cell division (22).

Studies that map initiation sites with high spatial resolution have revealed their frequent overlap with CpG islands (CGI) and transcription start sites (TSS; 4,5,23), as well as G-quadruplex (G4)-forming structures (24,25). While no consensus sequence defines origin activity in mammalian cells, many origins share a G-rich signature (7) and a short (~40 bp) region that is deficient in common single nucleotide polymorphism variants and indels (26).

Pioneering work by J.H. Taylor (27) described how cells artificially held in S phase increased the number of replication initiation sites. Since then, flexibility in origin usage has been reported in different systems, including Drosophila and mammalian cells (4,5,28,29). This flexibility is facilitated by the fact that initiator proteins ORC, CDC6, CDT1 and MCM2–7 'license' many more origins than those actually needed to duplicate the genome. A large number of origins apparently remain in a dormant state but have a chance to become active in situations of replicative stress (RS), when forks are slowed or stalled by DNA lesions, collisions with the transcriptional machinery or other factors (30,31). The activation of dormant origins provides a compensatory mechanism to complete genome duplication (32,33) and their relevance *in vivo* has been supported by the characterization of mouse strains hypomorphic for MCM2–7, which suffer from stem cell deficiencies, anemia and cancer (34). On the other hand, the availability of extra origins may pose a risk upon oncogenic stimuli that induce promiscuous origin activity and increase the frequency of collisions between replication and transcription forks (14). A better understanding of these processes requires in-depth information about the genomic positions and regulation of 'regular' versus stress-responsive origins.

Here, we have performed a comparative study of origin activity in mouse embryonic stem cells (mESCs) under conditions of mild RS that affects the frequency of activation of most initiation sites. The stratification of origins in activation efficiency groups and their integration into three-dimensional (3D) networks of chromatin interaction offer new insights about the elusive determinants of origin regulation in mammalian cells.

MATERIALS AND METHODS

Cell lines, culture and manipulations

TetO-CDC6 mESCs derived from TetO-CDC6 mice (35) were cultured on 0.1% gelatin-coated plates in Dulbecco's modified Eagle's medium (DMEM) with Ultraglutamine 1 and 4.5 g/l glucose (Lonza) supplemented with 15% FBS (Sigma), 50 U/ml penicillin–50 mg/ml streptomycin (Invitrogen), minimum essential medium non-essential aminoacids (MEM NEA; Invitrogen), 100 μM 2-mercaptoethanol (Invitrogen) and 10^3 U/ml ESGRO mLIF medium supplement (Millipore). To induce CDC6 overexpression, 1 μg/ml doxycycline (dox, Sigma) was added to the medium for 30 h. When indicated, mESCs were treated with 0.5 μM aphidicolin (Sigma-Aldrich) for 2.5 h to induce mild RS.

Flow cytometry

To monitor DNA content, cells were stained overnight with 50 μg/ml propidium iodide (PI; Sigma-Aldrich) in the presence of RNase A (10 μg/ml, Qiagen). In order to analyse DNA synthesis, cells were pulse-labelled with 20 μM BrdU for 30 min, trypsinized, washed in PBS and fixed with –20°C 70% ethanol for 24 h. 2 M HCl was added for 20 min at RT, before washing cells twice with PBS and incubating in blocking solution (1% bovine serum albumin in PBS, 0.05% Tween-20) for 15 min at RT. FITC-conjugated anti-BrdU antibody (BD Biosciences Pharmigen) was added for 1 h at 37°C. Samples were analyzed in a FACS Canto II cytometer (Becton-Dickinson) and data was processed with FlowJo V 9.4 or V.10.1 (Three Star).

Whole cell extract preparation and immunoblots

Cells were harvested and resuspended in Laemmli sample buffer (50 mM Tris–HCl pH 6.8, 10% glycerol, 3% SDS, 0.006% w/v bromophenol blue, 5% 2-mercaptoethanol) at 10^6 cells/ml. Extracts were sonicated for 30 s at 15% amplitude in a Branson Digital Sonifier. Standard protocols were used for SDS-polyacrylamide gel electrophoresis and immunoblotting. Primary antibodies used in this study are listed in Supplementary Table S5. Horseradish peroxidase (HRP)-conjugated secondary antibodies (GE Healthcare) and ECL developing reagent (Amersham Biosciencies) were used.

Analysis of DNA replication in stretched DNA fibers

Cells were pulse-labelled sequentially with 50 μM CldU (20 min) and 250 μM IdU (20 min), harvested and resuspended in PBS (0.5×10^6 cell/ml). 2 μl drops of cell suspension were placed on microscope slides and lysed with 0.5% SDS, 0.2 M Tris pH 7.4, 50 μM EDTA in 10 μl for 6 min at RT. Slides were tilted 15°C to spread DNA fibers, air-dried, fixed in –20°C methanol:acetic acid (3:1) for 2 min and stored at 4°C overnight. Slides were then incubated in 2.5 M HCl (30 min/RT) to denature DNA and washed (3×) in PBS. Blocking solution (1% bovine serum albumin in PBS, 0.1% Triton X-100) was added for 1 h at RT. Slides were incubated with anti-CldU, IdU and ssDNA primary antibodies

for 1 h at RT, washed and incubated with the corresponding secondary antibodies for 30 min. Prolong mounting media (Invitrogen) was used. Images were acquired in a DM6000 B Leica microscope with an HCX PL APO 40×, 0.75 NA objective. Fork rate values were derived from the length of IdU tracks, measured using ImageJ software, and a conversion factor of 1 μm = 2.59 kb (36). >300 tracks were measured per condition. The percentage of origins activated during the first pulse (green-red-green tracks) was quantified relative to all replicative structures containing red signal. >500 total structures were scored in each case. Three biological replicates of each experiment were performed.

Short nascent strand purification

For each assay, 10^8 growing cells were harvested and incubated for 15 min in lysis buffer (0.5% SDS, 50 mM Tris pH 8.0, 10 mM EDTA,10 mM NaCl). 100 μg/ml Proteinase K (Roche) was added and samples were incubated overnight at 37°C. DNA was isolated by standard phenol purification and EtOH precipitation, resuspended in TE (10 mM Tris pH 8.0, 1 mM EDTA) supplemented with 0.1 U/μl RNAse-OUT (Invitrogen), and stored at 4°C for at least 48 h. Following heat denaturation (100°C/10 min), DNA samples were loaded onto 5–20% sucrose gradients and fractionated according to size by centrifugation (SW-40Ti rotor; Beckman Coulter Optima L-100 XP; 20 h/ 78 000 rcf/20°C) as described (37). DNA from approximately 13 × 1-ml fractions was precipitated with ethanol and analysed in 1% alkaline agarose gel electrophoresis. Fractions 4–5, corresponding to DNA fragments of 300–1500 nucleotides, were selected. For each experiment, DNA samples were pooled from two gradients.

DNA samples were treated with 100 U of T4 polynucleotide kinase (PNK, Thermo Fisher) in the presence of 1 mM dATP (Roche) and 40 U of RNAseOUT (Thermo Scientific) for 30 min at 37°C. The PNK reaction was stopped by the addition of 6.25 μg proteinase K, 0.125% sarkosyl and 2.5 μM EDTA (30 min/37°C). Samples were heat-denatured (95°C/5 min) and incubated o/n at 37°C with 150 U of λ-exonuclease (λ-exo; Thermo Scientific) in the presence of 40 U of RNAseOUT. Reactions were heat-inactivated (75°C/10 min) and DNA was recovered by EtOH precipitation. To increase the purity of SNS, three cycles of PNK treatment and λ-exo digestion were performed. Each digestion step was controlled by adding 50 ng of linearized pFRT-myc plasmid (provided by Dr. Susan Gerbi, Brown University, USA) to 5% of the digestion reaction and incubated in the same conditions. pFRT-myc contains two G-quadruplex-forming sequences, reported to be digested less efficiently by λ-exo (38). Control λ-exo reactions were analysed in 1% agarose gels to confirm full digestion of the plasmid. To control for SNS enrichment at the Mecp2 origin, qPCR reactions were performed in duplicates using ABI Prism 7900HT Detection System (Applied Biosystems) and HotStarTaq DNA polymerase (Qiagen) according to manufacturer's instructions. Data was analysed in Applied Biosystems Software SDS v2.4. Primer sequences for Mecp2 origin and Mecp2 flanking region are indicated in Supplementary Table S6.

SNS-Seq library preparation and high-throughput sequencing

RNA primers were removed from SNS molecules with RNAse A/T1 Mix (Roche) for 60 min at 37°C. 100 μg/ml Proteinase K was added (30 min, 37°C) and DNA was extracted and precipitated. ssDNA was converted to dsDNA using 50 pmol of random hexamer primers phosphate (Roche) as described (39). Primer extension was performed by incubation with 10 mM dNTPs (Roche) and 5 U exo- Klenow Fragment (New England Biolabs) for 1 h at 37 C followed by incubation with 80 U of TaqDNA ligase (New England Biolabs; 50 C/30 min). DNA was extracted, precipitated and resuspended in TE. For the input sample, 4×10^7 mESCs were lysed in 1% SDS, 50 mM Tris–HCl pH 8.0, 10 mM EDTA (2×10^7 cell/ml) and sonicated in a Bioruptor device (Diagenode) for 25 min at 30 s intervals. DNA was extracted with phenol/chloroform, ethanol-precipitated and resuspended in $0.5 \times$ TE. DNA libraries were prepared at the Fundación Parque Científico de Madrid (FPCM) using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (New England Biolabs) and purified with Agencourt AMPure XP beads (Beckman Coulter). Each library was sequenced using single-end 75 bp reads ($120–140 \times 10^6$ reads per sample) in a NextSeq500 System (Illumina).

SNS-seq data analysis

The quality of sequencing reads was analysed with FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/). Short sequences, adaptor sequences and read duplicates were removed. Reads were analysed with the RUbioSeq pipeline v3.8. (40) using SAMtools v0.1.19 (41), Picard tools v1.107 (broadinstitute.github.io/picard/), MACS v2.0.10 (42) and the GRCm38/mm10 mouse reference genome. When MACS was used, peak calling was performed vs input. When the Picard algorithm (4) was used, genome segmentation was based on RT data that accurately matches the read coverage differences between segments. Common peaks were obtained using BedTools v2.23.0 (43) with parameters: -f 0.1 -r -wa –u. For common peaks between MACS and Picard the genomic coordinates defined by Picard were used in additional analyses. Peaks located on chromosome Y were excluded from the analysis. The SNS-seq WT I dataset, taken from (6), was generated in parallel to all other SNS-seq samples, which were specifically designed for this study. Read distribution around peak centers was generated using seqMINER v1.3.3e. Low-mappability regions were identified using the scan quantile peak-calling algorithm (4) on a sequenced genomic DNA input, with the same parameters set for the SNS-seq samples. The resulting peaks, as well as the small gaps present in the genome segmentation needed for peak-calling, were marked as non-mappable regions (ShadeAreas files at github.com/VeraPancaldiLab/RepOri3D). Subtelomeric and pericentromeric regions from the UCSC database telomere annotation were extended by visual inspection in the browser and added to the list of non-mappable regions. Origin clustering was done using SeqMINER (44) for each condition independently, using both replicates for calculating the clusters. Default options were used except

for the number of clusters, which was empirically set to 6 as it gave clusters with comparable profiles in the three conditions.

Correlation with epigenomic features and chromatin states

Genomic coordinates of origins were converted from mm10 to mm9 genome assembly with LiftOver (https://genome.ucsc.edu/util.html). Origins were intersected with the genomic features indicated in Supplementary Table S7 or with a set of previously compiled epigenomic features (45). Each epigenomic feature dataset was 'discretized' in 200 bp windows: the presence of a given mark within a 200 bp window was scored as 1, and its absence as 0. The overlap between origin fragments and the genomic windows was calculated using findOverlaps in the GenomicRanges R package. The number of origins overlapping with each feature was calculated for the experimental sets of origins and for 1,000 sets of origins randomly shuffled along the genome (excluding low-mappability regions). The enrichment of origins at any particular feature was calculated as the ratio between the number of origins overlapping the feature and the median of all randomizations, and significance was calculated with empirical *P*-values. In the calculations of origin enrichment at chromatin states, some of the 20 states defined in (45) corresponding to similar chromatin functions were merged according to the indicated definitions (Supplementary Table S8). Overlap of origins with initiation zones and initiation sites from (5) was calculated using the tables in GEO GSE68347. Replication timing was assigned to replication origins based on their overlap with available RT data in mESCs (46).

Analysis of origin efficiency

In each sample, the efficiency of each individual origin was determined following three steps: (i) the read coverage in each origin, defined as the sum of reads covering each nucleotide (i.e. the sum of per base read depths) of the origin, was calculated using samtools function bedcov with default parameters. This analysis was done after downsampling every dataset to the lowest coverage obtained (APH-I). Down-samplings were performed a total of 10 times, and the median coverage was determined for each origin; (ii) background was calculated for each sample as the coverage of a similar number of randomized genomic fragments of the same length as 'called' origins, but forced not to overlap with them. After 100 such randomizations, a function of median background noise for each origin length was calculated. The estimated background noise for each origin was subtracted from the initial values; (iii) background-corrected values were normalized dividing by the length of the origin. Finally, mean efficiencies of origins from biological replicates were calculated as an average of efficiency between the two replicates. To calculate the distribution of origins across efficiency quantiles, the efficiency of all merged origins was calculated and split into deciles. Next, the proportion of origins falling into each quantile was calculated separately for common or responsive origins.

Simulated origin sets

Simulated sets of origins were obtained by relocating all origins in a different position choosing from the whole genome, excluding low-mappability regions. For the TSS distance-preserving simulated sets used in the network analysis, origins were placed at random genomic positions but maintaining the same distance from a TSS as the real ones. In this process, some candidate random origins were placed at new locations with the correct distance from the target TSS, but accidentally closer to another TSS. These candidate random origins were discarded and randomized again. If after 1000 randomization attempts, some origins (always <2%) still failed to match the randomization criteria, the distance from the TSS was progressively increased until these origins were successfully relocated. Randomizations preserving TSS and gene expression levels were done similarly to the randomizations preserving TSS, but allowing the origins to be located only close to genes with a similar gene expression, calculated using mESC data from datasets GSM2533843 and GSM2533844 (47). Genes were classified by their mean expression levels into 5 groups by cutting at 0.1, 1, 5 and 20 fpkm, and origins were relocated randomly close to a random gene of the group of genes with similar expression, preserving the distance to the closest TSS. Custom scripts can be found at https://github.com/VeraPancaldiLab/RepOri3D. Ten simulated origin sets were generated from each WT dataset replicate, for a total of 20 simulated sets.

Origin integration with chromatin interaction maps

The Virtual Origin Capture Hi-C network was generated from high depth Hi-C data (47). Bins of the Hi-C contact matrix at 5 kb resolution that contained WT origins were identified. All contacts involving these bins with a score >25 were extracted to generate a network of origin-origin contacts. The average efficiency over the Hi-C fragment containing the origin was calculated in each dataset (WT, APH, CDC6). In order to derive A- and B-compartment subnetworks, network nodes were assigned to A/B chromatin compartments according to (47). The Virtual Promoter Capture Hi-C was generated from the same Hi-C contact map by extracting 5 kb bins that overlapped with annotated TSSs. The average efficiency of each promoter-containing bin was calculated. The Promoter-Capture Hi-C (48) contact map was processed using the CHiCAGO pipeline, which identifies significant 3D contacts starting from the raw Capture Hi-C data (49). To define a TAD network, edges were established between pairs of TADs (47) when both included origins that interacted with each other. The efficiency value for each TAD was calculated as the average efficiency of the origins included in it.

Network analysis and RT data

All network correlative analyses between degree, efficiency and RT were performed using the igraph package in R (www.igraph.org). Efficiency and RT features were mapped onto the contact networks using ChAseR as described (50).

TAD definitions were taken from (47). Networks were visualised using Cytoscape. RT data for mESC cell lines was downloaded from https://www2.replicationdomain.com/ (Accession Int52769503).

Assortativity analysis and randomizations

Assortativity of Origin Efficiency (OriEfAs) was estimated using Chromatin Assortativity (ChAs; (51)) as calculated by the ChAseR package (50). OriEfAs is defined as the Pearson correlation coefficient of the efficiency of an origin across all pairs of nodes connected with each other in a specific network. ChAs for a particular feature is analyzed in relation to the abundance of said feature. For example, if a particular mark is found in the majority of the fragments in the network, its localization in specific areas of the network cannot be observed and the value of ChAs will be low. On the contrary, if a certain feature is detected only in a small subset of fragments, but these fragments interact preferentially with each other, ChAs will be high. For this reason ChAs should be evaluated against an expected value that can be obtained by randomizing the association of each node to a specific feature value, while preserving the frequency distribution of the features across the nodes in the randomizations. When domains of a specific feature span more than one fragment along the chromosome, high values of ChAs may only reflect the higher probability of 3D contact for regions that are close in 1D. Using the ChAseR randomization strategy, which preserves the distribution of distances spanned in the network contacts, we can estimate how much the ChAs value is dependent on the contacts between non-adjacent regions along the genome. *Z*-scores are calculated as the number of standard deviations between the experimental value and the mean of the randomizations, estimating the significance of the ChAs values observed as compared to expectations based on purely 1D correlations.

Statistical analysis

Statistical analyses relative to the wet-lab experiments were performed using Prism v4.0 (GraphPad Software) or Microsoft Excel v15.38. For comparison of two data groups, two-tailed paired Student's t-test was used. In the analysis of fork rate in stretched DNA fibers, a nonparametric Mann–Whitney rank sum test was used. To statistically assess the overlap between origins and genomic features (or chromatin marks), we compared the number of overlaps between origins and features to the number of overlaps calculated after random re-localization of the origins in the genome (1000 randomizations). Empirical p-values were obtained as $(r + 1)/(n + 1)$, where n is the number of randomizations and r is the number of randomizations that produce a test statistic greater than or equal to that calculated for the real data. Statistical analyses for the computational part were performed using R. Scripts and R notebooks can be found at https://github.com/VeraPancaldiLab/RepOri3D.

RESULTS

Mapping mESC replication origins under stress

To identify replication initiation sites in mESCs with high resolution, we used deep sequencing of short nascent strands (SNS-Seq; Supplementary Figure S1A), a method that has yielded reproducible results in many laboratories. SNS-Seq was performed in normal cell growth conditions (hereafter referred to as WT) or in two experimental settings that enhance origin activity: (i) exposure to aphidicolin (APH), a DNA polymerase inhibitor; (ii) ectopic expression of CDC6, a limiting factor for origin licensing and activation (35,52,53). APH induces RS by slowing down replication forks and triggering the activation of extra origins as a compensatory mechanism. In turn, CDC6 overexpression aims at enhancing origin activity directly, with subsequent fork slowdown caused by reduced dNTP availability (Figure 1A; (54)). In the experimental conditions used, both APH treatment and CDC6 overexpression increased origin usage and slowed down forks without preventing overall DNA synthesis, inducing extensive DNA damage or causing detectable DNA re-replication (Supplementary Figure S1B–D). Two replicates of the SNS-Seq assay were performed for each condition. The quality of SNS preparations was monitored by controlling the completeness of lambda-exonuclease (λ-exo) digestion, a necessary step to eliminate false positives caused by broken DNA ((37); Supplementary Figure S1E), and by confirming the enrichment at a known origin relative to its flanking region ((23); Supplementary Figure S1F).

Two independent methods were used to analyze SNS profiles: MACS, a ChIP-Seq tool that has been applied before to SNS-Seq data (29), and a dedicated algorithm optimized for SNS-Seq (4) that also takes into account local coverage heterogeneities (Figure 1A-B). The total number of peaks was in the range of 71 435–94 747 for MACS and 41 376–94 893 with the Picard algorithm (Supplementary Table S1). A visual inspection of the genome browser revealed essentially flat profiles of SNS-Seq reads in sonicated genomic DNA or in a DNA sample treated with RNAse prior to l-exo digestion, generated in a previous work ((5); Figure 1B, last two rows). Reproducibility between biological replicates was assessed by pairwise correlation of origin number per genomic segment, and by the distribution of SNS-seq reads for both replicates around the peak centers of one of them (Supplementary Figure S1G). To minimize the influence of technical variability in subsequent analyses, only those peaks called by both algorithms in the two replicate assays were included in the origin datasets. With these stringent criteria, 20 174; 31 685 and 31 402 active origins were defined in WT, APH and CDC6 conditions, respectively. Their characteristics in terms of size, replication timing and overlap with previous origin maps defined by SNS-seq in mESCs (5) are summarized in Supplementary Table S2.

Genetic and epigenetic features of mESC replication origins

In the three datasets (WT, APH and CDC6), initiation sites were found both within gene bodies and at intergenic locations. However, their overlap with CpG islands (CGIs), transcriptional start sites (TSS) and exons occurred with much higher frequency than expected if origins were randomly distributed along the genome, in agreement with previous studies (Figure 1C, top). Origin localization analysis at 'chromatin states' defined by specific combinations of epigenetic features (45,55) revealed a high enrichment at en-
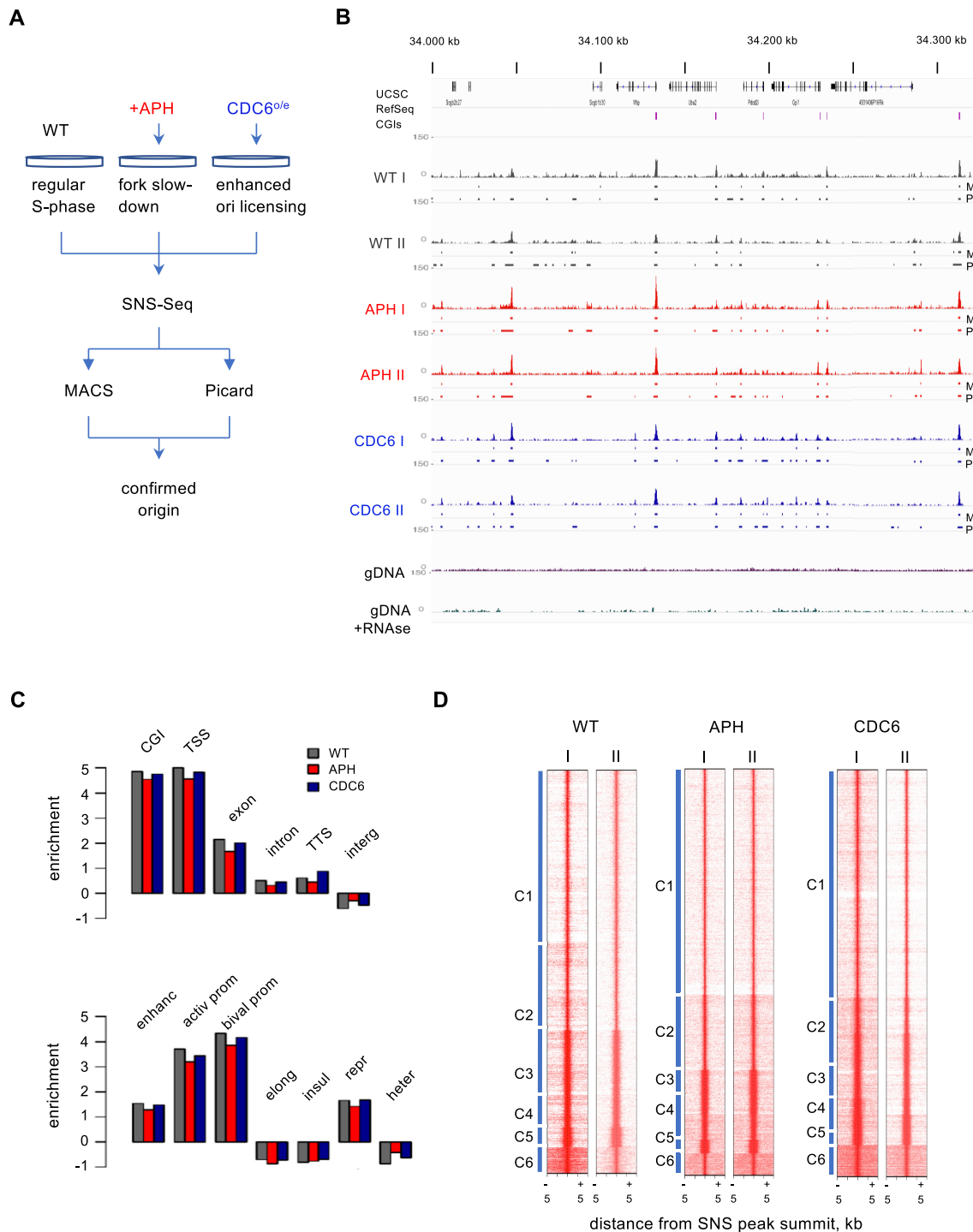
Figure 1. Genome-wide mapping and features of mESC replication origins in normal and stress conditions. (**A**) Schematic of the experimental approach. (**B**) Genome browser image showing read density tracks in a representative fragment of chromosome 19 from two SNS-Seq replicates in control mESCs (WT), mESCs treated with aphidicolin (APH) and mESCs after CDC6 overexpression (CDC6). Vertical dashes indicate the positions of peaks called by MACS (M) and Picard (P) algorithms. Bottom tracks correspond to SNS-seq reads from sonicated genomic DNA (gDNA) and a control DNA sample treated with RNAse before l-exo digestion from (5). Sequencing coverage for the control gDNA was approximately 2.5× of the mm10 genome annotation. See also Supplementary Figure S1. (**C**) Enrichment of WT, APH, and CDC6 origins at the indicated genomic features (top panel) or chromatin states (bottom panel), relative to randomized controls. Enrichments were calculated as the $\log_2$ ratio between observed and expected origins overlapping with a feature, the expected value being the median of overlapping origins across randomizations. All enrichments were significant at $P < 0.001$ (permutation test). CGI, CpG island; TSS, transcription start site; TTS, transcription termination site. Other abbreviations: interg(enic), enhanc(er), activ(e) or bival(ent) prom(oter); elong(ation), insul(ator), repr(essed), (heter)ochromatin. More detailed analyses are shown in Supplementary Figure S2. (**D**) Heat maps showing the distribution of WT, APH or CDC6 SNS-seq reads from experimental replicates I and II within a 10 kb window centered at the origin summit defined by the Picard algorithm. Origins at each dataset were classified in 6 groups by unsupervised clustering analysis of SNS reads distribution (C1–C6). See also Supplementary Figure S3.

hancers, promoters (active and bivalent) and the Polycomb-repressed state. In contrast, heterochromatin and the 'transcriptional elongation' state contained origins with lower frequency than expected from a random origin distribution (Figure 1C, bottom). The co-localization of origin sets with individual chromatin proteins is shown in Supplementary Figure S2.

The structural patterns of SNS reads and their associations with genomic features have been used to stratify mESC origins in different classes (5). Unsupervised clustering analysis of read density distribution around peak centers (10 Kb window) defined six comparable origin classes at WT, APH and CDC6 datasets (C1-C6; Figure 1D, Supplementary Figure S3A, Table S3). C1, the largest one, predominantly corresponds to intergenic origins with a modest enrichment in specific epigenetic signatures (Supplementary Figure S3B). This is in agreement with the previous finding that > 40% of mESC origins are isolated peaks with low density of SNS reads (5). Classes C3–C5 displayed higher levels of SNS reads per origin and were enriched in genomic locations associated with transcriptional initiation. C2 comprises an intermediate category between C1 and C3–C5, in terms of their enrichment at SNS and epigenetic signatures, and C6 includes more diffuse initiation sites and were mainly enriched at enhancer elements. While overall origin clustering was similar in the three datasets, SNS abundance was increased in APH or CDC6 datasets relative to WT, particularly in classes C2, C4 and C5 (Supplementary Figure S3C).

Stress-responsive origins are also active in the unchallenged S phase

The intersections between WT, APH and CDC6 datasets revealed a subset of almost 12 000 origins identified in all conditions, which we termed 'common' (COMM). Two large (>17,000) groups of origins were apparently responsive to aphidicolin (APH-R) or CDC6 (CDC6-R), with almost half of them (>8200) responding to both stimuli (APH + CDC6-R; Figure 2A). However, when sequencing reads were inspected in a genome browser, the majority of positions corresponding to responsive origins also displayed some SNS enrichment in WT cells, and many of them were scored as active origins by one of the two peak-calling algorithms in at least one of the WT replicates (Figure 2B). Heatmap representation of WT SNS-Seq reads around APH-R or CDC6-R peak centers confirmed that APH- or CDC6-responsive origins were partially active in untreated conditions (Figure 2C). Responsive origins tend to be shorter in terms of peak width and later-replicating than COMM origins (Supplementary Table S2), suggesting different genomic characteristics that regulate their activation.

The activation efficiency of origins is proportional to the density of SNS reads in each peak (normalized SNS-Seq counts by peak length; (28,29,56)). When origins were classified in 10 quantiles, from the least active (Q1) to the most active (Q10); Figure 2D), the majority (>70%) of COMM origins were included in the three most-efficient quantiles, resembling the distribution of initiation events reported in non-transformed human cells (7). In contrast, origins in the APH-R, CDC6-R and APH + CDC6-R subsets displayed

a broader distribution that covered the whole range of efficiencies (Figure 2E). An alternative estimation of activation efficiency, based on total SNS counts relative to sequencing coverage, yielded similar results (Supplementary Figure S4A, B), ruling out possible biases due to differences in peak size. Combined, these results suggest that mESC COMM origins correspond to higher-efficiency initiation sites irrespectively of RS, whereas responsive origins correspond to initiation sites already used in the unchallenged S phase, which become active in a higher proportion of cells upon stress. In agreement with this interpretation, efficient COMM origins co-localize with CGIs, TSS, exons, enhancers, active and poised promoters with higher frequency than stress-responsive origins (Figure 2F).

RS increases the frequency of origin activation

To investigate how the frequency of activation of individual origins is regulated, we first checked the genetic and epigenetic signatures of WT origins stratified according to their efficiency (for graphical simplicity, only four efficiency quartiles were used). As anticipated in earlier studies (23), direct correlations were observed between origin efficiency and features of active transcription and chromatin (Figure 3A). Origin efficiency strongly correlated with the proximity to the nearest TSS (Figure 3B), and this effect becomes more pronounced for TSSs driving higher levels of gene expression (Figure 3C). Direct correlations were also observed between origin efficiency and several epigenetic marks, including activating histone modifications H3K9ac, H3K4me as well as histone variant H2A.Z (Supplementary Figure S4C). These analyses underscore the predictive value of active transcription initiation sites for origin specification and support the notion that the open chromatin environment associated with transcriptional competence contributes to the recruitment of origin-activating proteins.

We next measured the extent of initiation activity at each origin in the different experimental conditions by pairwise comparisons of the averaged ratios of normalized sequencing reads per peak. When comparing origin efficiency in APH versus WT, or CDC6 versus WT conditions, the distribution of origins deviated from the diagonal slope, indicating that most origins increase their activity upon RS. Furthermore, density plots showed that the vast majority of origins categorized as responsive were indeed preferentially activated under RS conditions (colored dots in Figure 3D and E; 95% and 93% of APH-R and CDC6-R origins located above the diagonal, respectively). In these pairwise comparisons, some efficiency differences were observed between the two WT replicates, and to a lesser extent, between the two CDC6 replicates, which we attribute to experimental variability (Supplementary Figure S5A). Despite this fact, the increase in efficiency in APH-R and CDC6-R origins was unequivocally observed when they were compared against each one of the individual WT replicates (Supplementary Figure S5B and C).

Consistent with the distribution shown in Figure 2E, the range of efficiency values of COMM origins (calculated for each origin as the average of the three experimental conditions) was higher than that of APH-R and CDC6-R origins (Figure 3F). Indeed, the median efficiency of COMM ori-
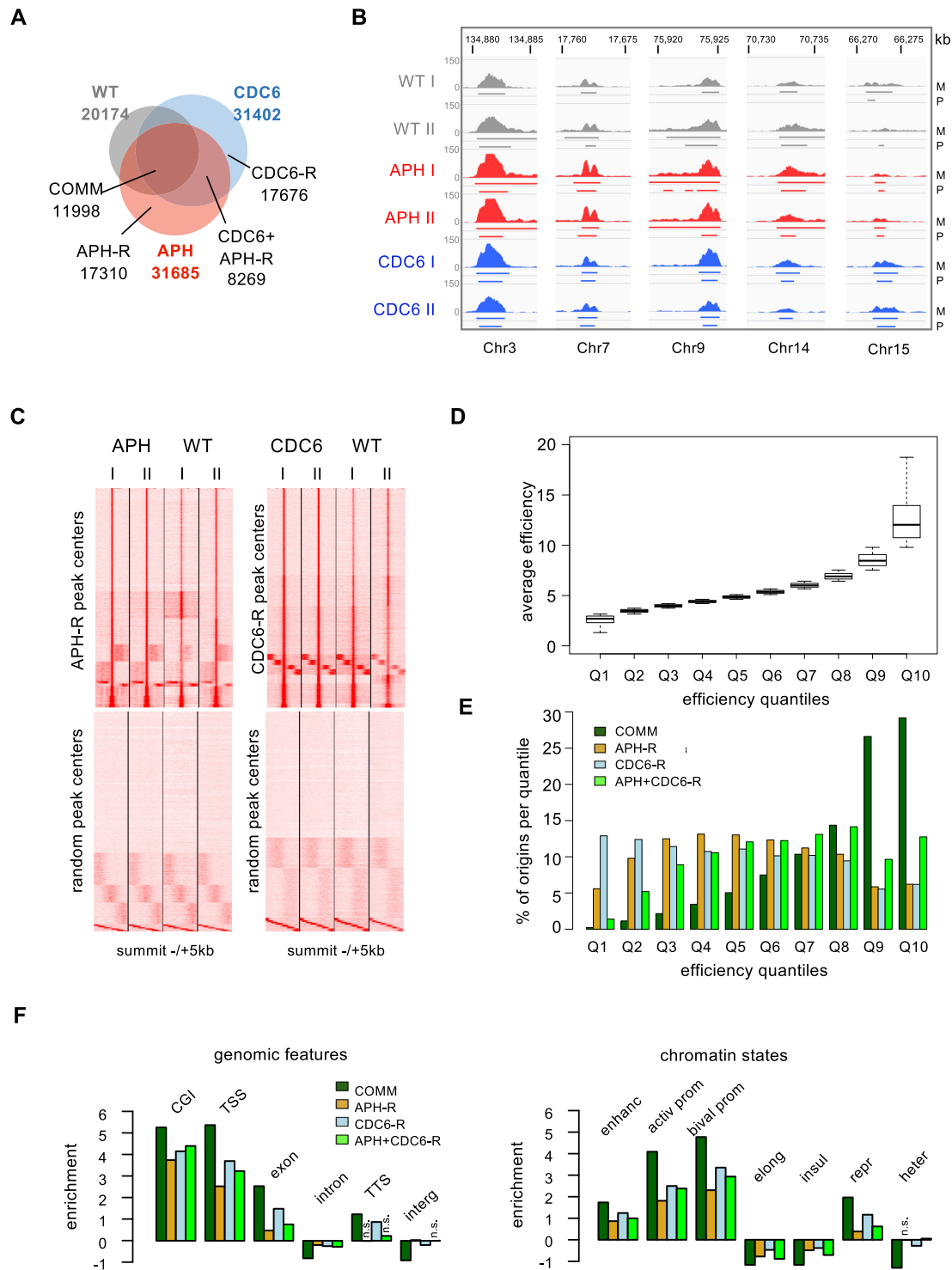
Figure 2. Stress-responsive origins are also active in control conditions. (**A**) Venn diagram of origin subsets, determined by intersections of WT, APH and CDC6 datasets. (**B**) Genome browser examples of stress-responsive origins. (**C**) Heatmap representation of the distribution of SNS-seq reads for the indicated experimental replicates around APH-R, CDC6-R or randomised peak centres (bottom panels). (**D**) Box-plot showing the average efficiency, divided into 10 quantiles of merged WT, APH and CDC6 origins, ranging from the 10% with lowest averaged efficiency (Q1) to the 10% with highest averaged efficiency (Q10). The horizontal line within each boxplot represents the median, whereas the lower bound of the box defines the first quartile and the upper bound of the box defines the third quartile. Bottom and top of the whiskers represent minimum and maximum numbers respectively for each boxplot excluding outliers, which are not represented. See also Supplementary Figure S4A for a similar analysis applying a different method to calculate origin efficiency. (**E**) Distribution of COMM, APH-R, CDC6-R and APH + CDC6-R origins across the efficiency quantiles. See also Supplementary Figure S4B. (**F**) Enrichment of COMM and responsive origins at the indicated genomic features (left) and chromatin states (right), relative to randomized controls. Enrichments and abbreviated labels are as in Figure 1C. All enrichments were significant at $P < 0.001$ unless indicated with n.s. (not significant, permutation test).
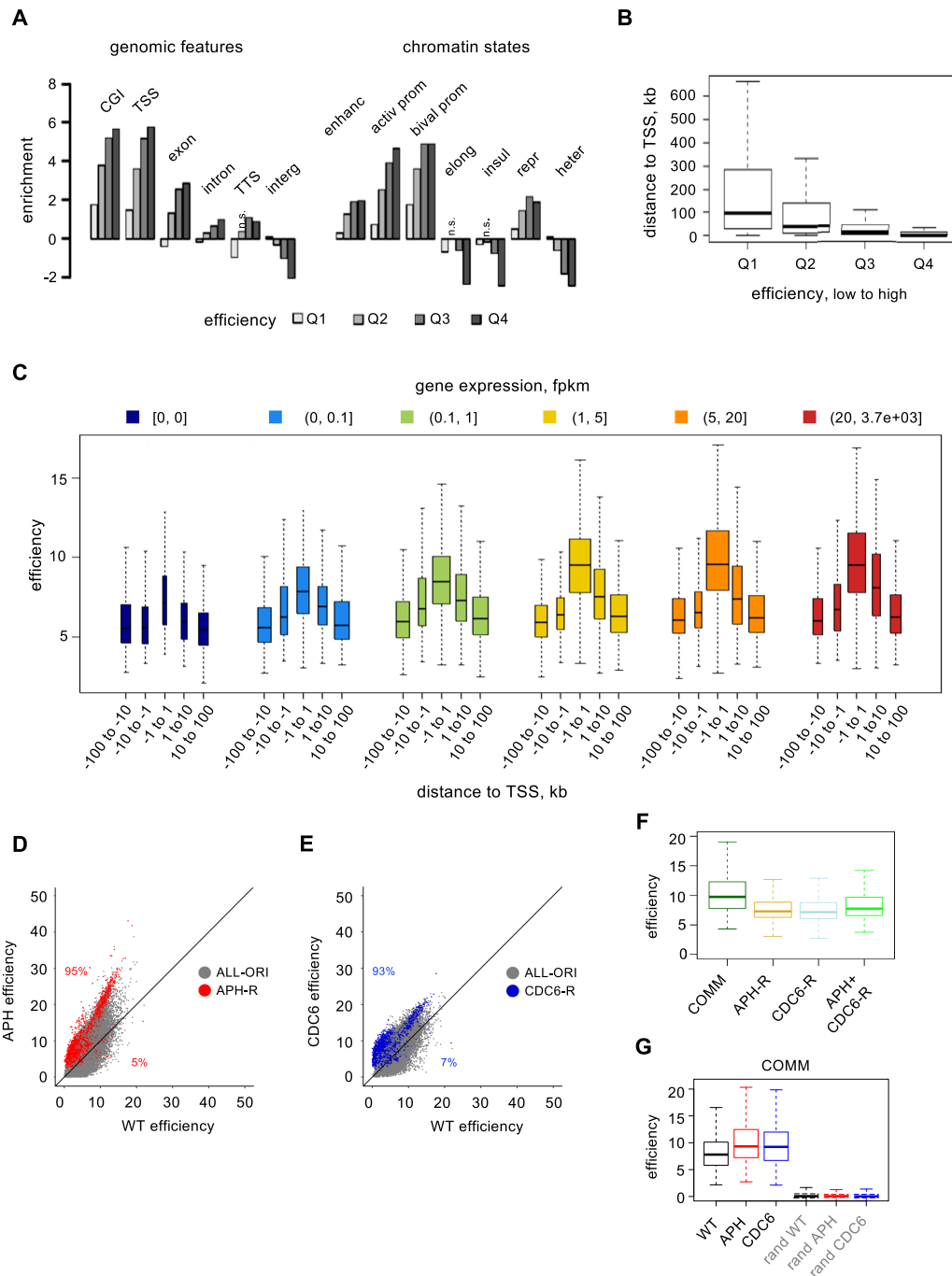
Figure 3. Increased activation of pre-existing origins upon stress. (**A**) Enrichment of WT origins, distributed in quartiles according to their efficiencies, with genomic features (left) and chromatin states (right), relative to randomized controls. Enrichment calculations and abbreviated labels are as in Figure 1C. Quartile efficiency values: Q1 = [2.1–5.1], Q2 = (5.1–6.5], Q3 = (6.5–8.8], Q4 = (8.8–29.3]. All enrichments were significant at $P < 0.001$ unless indicated with n.s. (not significant, permutation test). (**B**) Box plots showing the distances to the nearest TSS of efficiency-stratified WT origin datasets. Quartile efficiencies are the same as in (A). A significant negative correlation between origin efficiency and distance to TSS was observed. Pearson's correlation coefficient $r = 0.24$, *P*-value 6.8e-258. (**C**) Box plots showing the efficiency of WT origins grouped according to their distance to the closest TSS for six different gene expression levels (fpkm: fragments per kb of transcript per million mapped reads). Pearson's correlation coefficient and *P*-values as follows: $r = -0.13$, $P = 5.21e-06$ in [0,0]; $r = -0.15$, $P = 1.56e-21$ in (0,0.1]; $r = -0.28$, $P = 3.37e-72$ in (0.1,1]; $r = -0.28$, $P = 1.36e-79$ in (1,5]; $r = -0.28$, $P = 2.62e-78$ in (5,20]; $r = -0.31$, $P = 1.87e-54$ in (20,3.7e + 03]. The width of the box is proportional to the square-root of the number of observations in the groups. (**D**) Density plots comparing origin efficiency in APH versus WT conditions. APH-R origins are highlighted in red. The percentages of responsive origins located above or below the diagonal are indicated. (**E**) Same as (D), comparing origin efficiency in CDC6 versus WT conditions. CDC6-R origins highlighted in blue. See also Supplementary Figure S5B and C for pairwise comparisons of APH or CDC6 versus individual WT replicates. (**F**) Box plots showing the efficiencies of common and stress-responsive origins. Statistical significance (Student's *t*-test): COMM versus APH-R, $P = 3.7e-86$; COMM versus CDC6-R, $P = 9.0e-307$; COMM versus APH + CDC6-R, $P = 0.26$ (n.s.). (**G**) Efficiencies of common origins in WT, APH and CDC6 datasets. The efficiencies of randomized equivalent sets are also shown (dashed boxes). Statistical significance (Student's *t*-test): all comparisons (WT versus APH, WT versus CDC6, WT versus any randomization) were statistically significant with $P < 10e-16$.

gins was higher in APH or CDC6 conditions relative to WT cells (1.2-fold and 1.18-fold, respectively; differences significant at $P < 10^{-16}$; Figure 3G). Conversely, origins located at responsive positions displayed lower efficiency in WT cells than upon APH or CDC6 stimuli (Supplementary Figure S5D and E). No differences in efficiency were observed in equivalent sets of randomized genomic positions mimicking origins (Supplementary Figure S5D and E). These analyses confirm that responsive origins correspond to actual initiation sites that become activated with higher frequency under stress conditions.

Chromatin connectivity correlates with origin efficiency and RT

The execution of the temporal program of DNA replication is intrinsically connected to the 3D organization of the genome (57). To provide three-dimensional context, the origin datasets defined in this study were integrated into chromatin contact maps represented as networks in which chromatin fragments are 'nodes' and experimentally-determined interactions between them are represented as 'edges' (Figure 4A–D; (58–61)). We selected the most comprehensive Hi-C dataset available for mESCs (47) and derived a virtual origin-capture Hi-C network (VOCHi-C) by extracting those fragments that contained at least one origin in the WT dataset (Figure 4A). The nodes of VOCHi-C were defined as 5 kb regions surrounding the origin midpoint and their interaction scores were extracted from the Hi-C matrix. VOCHi-C contained 20 067 nodes, including >99% of the experimentally identified WT origins, and 786 101 connections between them (see Supplementary File 1 for network details). Increasing the window size of origin-containing fragments to 10 or 25 kb did not change the number of nodes, but increased the number of edges 1.5- and 2.3-fold, respectively.

To investigate the relationship between origin efficiency and VOCHi-C network properties, averaged efficiencies between WT replicates were assigned to each origin-containing node (Figure 4E; Materials and Methods). The 5 kb-resolution VOCHi-C network from chromosome 1 with the corresponding efficiency values is shown in Figure 4F. Note that this graphical representation does not indicate the actual positions of chromatin fragments within the nucleus, but the network of interactions established between origin-containing fragments in the cell population. This representation illustrates that most of the origins reside in large connected components linked by multiple interactions, rather than forming small separated clusters. A similar organization was found in all chromosomes (not shown). Visual examination of this network (Figure 4F, inset) suggested that larger nodes (i.e. those that establish more connections) displayed higher efficiency. In network theory, the number of connections established by a node with other nodes is referred to as 'degree'. A strong positive correlation was observed between node degree and origin efficiency in the whole VOCHi-C network (Figure 4G). This result indicates that origins located at more connected nodes (hubs) are activated more frequently in the population. This correlation was also detected when the nodes containing promot-

ers were depleted from the VOCHi-C network (Figure 4H), indicating that the higher efficiency of origin hubs is not solely explained by the higher connectivity of promoters in the network. Similar analyses increasing the window size of chromatin-containing origins yielded comparable results (Supplementary Figure S6A and B). As expected, highly efficient COMM origins displayed higher degree than the rest of WT origins (Figure 4I).

Higher origin density and efficiency has been reported in early-replicating chromosomal domains (5,7,28,39). Cross-analyses of WT origins stratified by efficiency with available mESC RT data (46) confirmed that more efficient origins tend to be early-replicating (Supplementary Figure S7A and B). Therefore, a positive correlation between origin connectivity and RT can be predicted. Indeed, origins located at nodes with higher degree displayed earlier RT than the rest (Supplementary Figure S7C). Taken together, these results indicate that origins at highly connected nodes tend to activate early in S-phase and display higher efficiency in the cell population.

Origin efficiency is assortative in chromatin interaction networks

Representing chromatin contacts as networks enables the use of specific analysis tools such as chromatin assortativity (ChAs) to study spatial patterns of interacting origins within the network (50,51). ChAs is a correlation coefficient, ranging between –1 and 1, measuring the extent to which the value of a given feature of a chromatin fragment is correlated with the values of the same feature in the fragments that interact with it. ChAs thus indicates the presence of preferential contacts between chromatin fragments with similar values of a specific feature (Figure 5A). In the context of this study, assortativity of origin efficiency (OriEfAs) can be used to reveal the tendency of origins located in interacting chromatin fragments to fire with similar efficiencies. Positive OriEfAs values were detected for WT, CDC6, APH and COMM origins in the VOCHi-C network (full small circles in Figure 5B). In each case, OriEfAs values (0.09–0.14 range) were significantly higher than OriEfAs values produced by computational randomization of the networks that preserved the distributions of linear distances between interacting fragments and the number of origins (empty large circles; range 0.031–0.035 for WT). Z-scores, used to estimate OriEfAs significance (Methods), were in the 69–117 range. In addition, OriEfAs was almost negligible in simulated sets of origin positions, even when the distance to the nearest TSS and the level of expression of the corresponding gene were preserved (full purple circle, overlapping with larger empty purple circles; OriEfAs range 0.006–0.013; Z-score ~8). These results indicate that origins that interact with each other have a statistically significant tendency to be activated with similar efficiency.

Given the strong correlation between the efficiency of origins and their proximity to TSSs (Figure 3A–C), OriEfAs was also tested in a Virtual Promoter-Capture Hi-C (VPCHi-C) network generated by selecting chromatin contacts between annotated TSS regions of the genome.
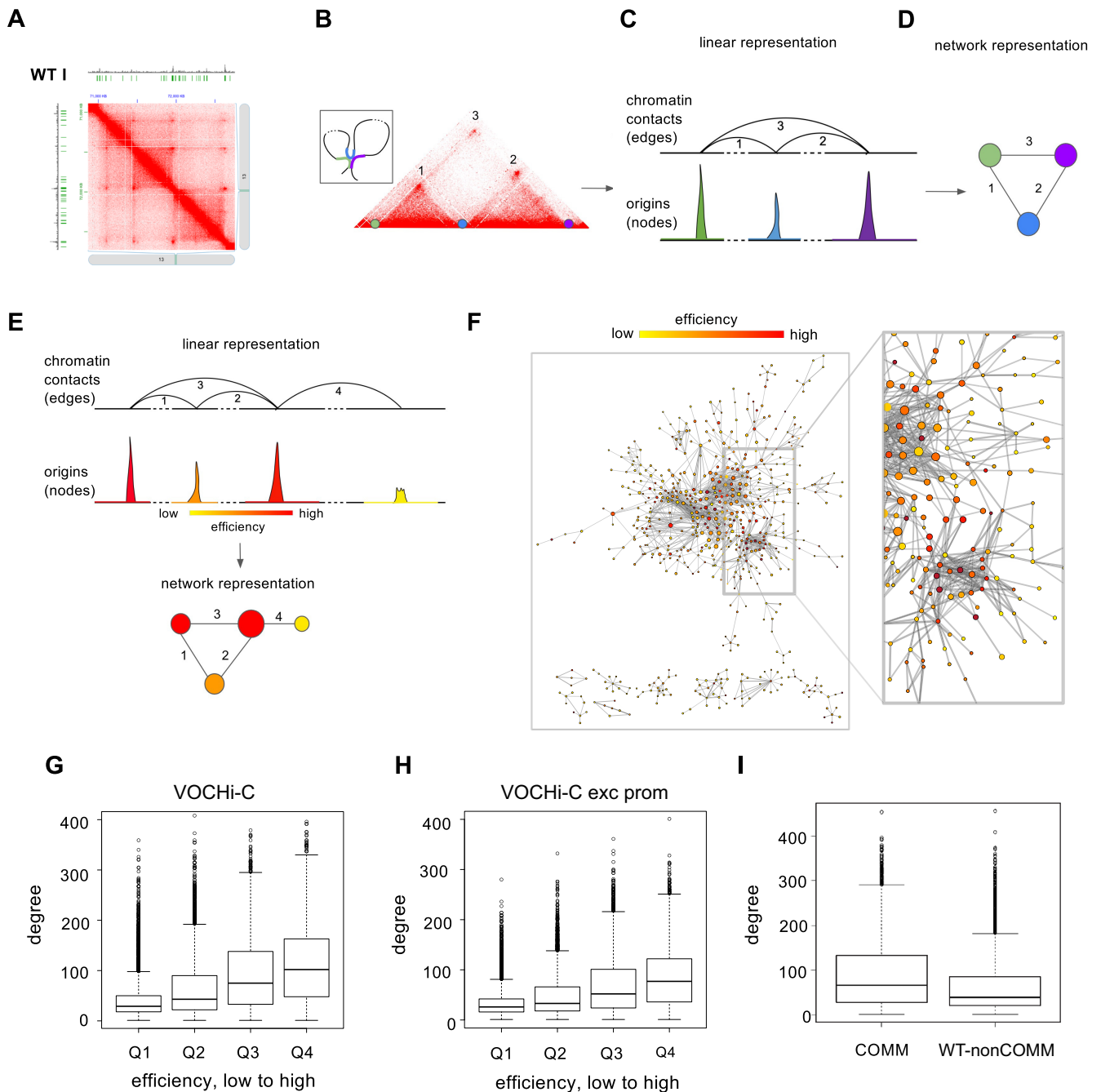
Figure 4. A Virtual Origin Capture Hi-C network (VOCHi-C). (**A**) Juicebox browser view of Hi-C contact map at 5 kb resolution for ESCs (47) of a representative region in chromosome 13. WT-I SNS-seq track and the corresponding peaks called by MACS algorithm are shown. (**B**) Schematic showing the identification of origins and extraction of their contacts from Hi-C data. Chromatin fragments containing origins (nodes) were coloured in green, blue and purple. Contacts (edges) were numbered to facilitate the location of a particular interaction between two origins. The inset shows possible chromatin folding of the Hi-C region in the example. (**C**) Representation of identified contacts between origins (top) and SNS-seq read peaks at the corresponding origins (bottom). (**D**) Resulting network representation: nodes are origins and edges (connections) indicate that origins reside in fragments that contact each other in the Hi-C data. (**E**) Schematic showing the integration of Hi-C chromatin contact data with replication origin efficiency signal and their representation in a network. Node size is proportional to the number of contacts with other nodes, while node colour reflects the efficiency value. (**F**) Illustration of VOCHi-C network (chromosome 1; 5 kb resolution), prepared as indicated in (E). For visualization purposes, only connections with score >50 are shown. In the left panel, all network edges are represented in grey lines. In the right inset, the edges stablished by a single, highly-connected node are indicated in black. (**G**) Correlation between node degree and origin efficiency in the whole VOCHi-C network. WT origins were stratified in quartiles according to their efficiencies: Q1 = [0.0–0.9], Q2 = (0.9–1.6], Q3 = (1.6–2.7], Q4 = (2.7–12.6]. Pearson's $R = 0.35$, $P < 10^{-16}$. See also Supplementary Figure S6. (**H**) Correlation of node degree and WT origins efficiency in a promoter-depleted network (VOCHi-C exc prom). Origins were stratified in quartiles as in (G). Pearson's $R = 0.33$, $P < 10^{-16}$. (**I**) Boxplot showing the distribution of node degrees in COMM and WT-non common origins. The mean number of connections of each COMM origin is 1.41-fold higher than the number of connections of each WT-non common origin ($P < 10^{-16}$ Wilcoxon).
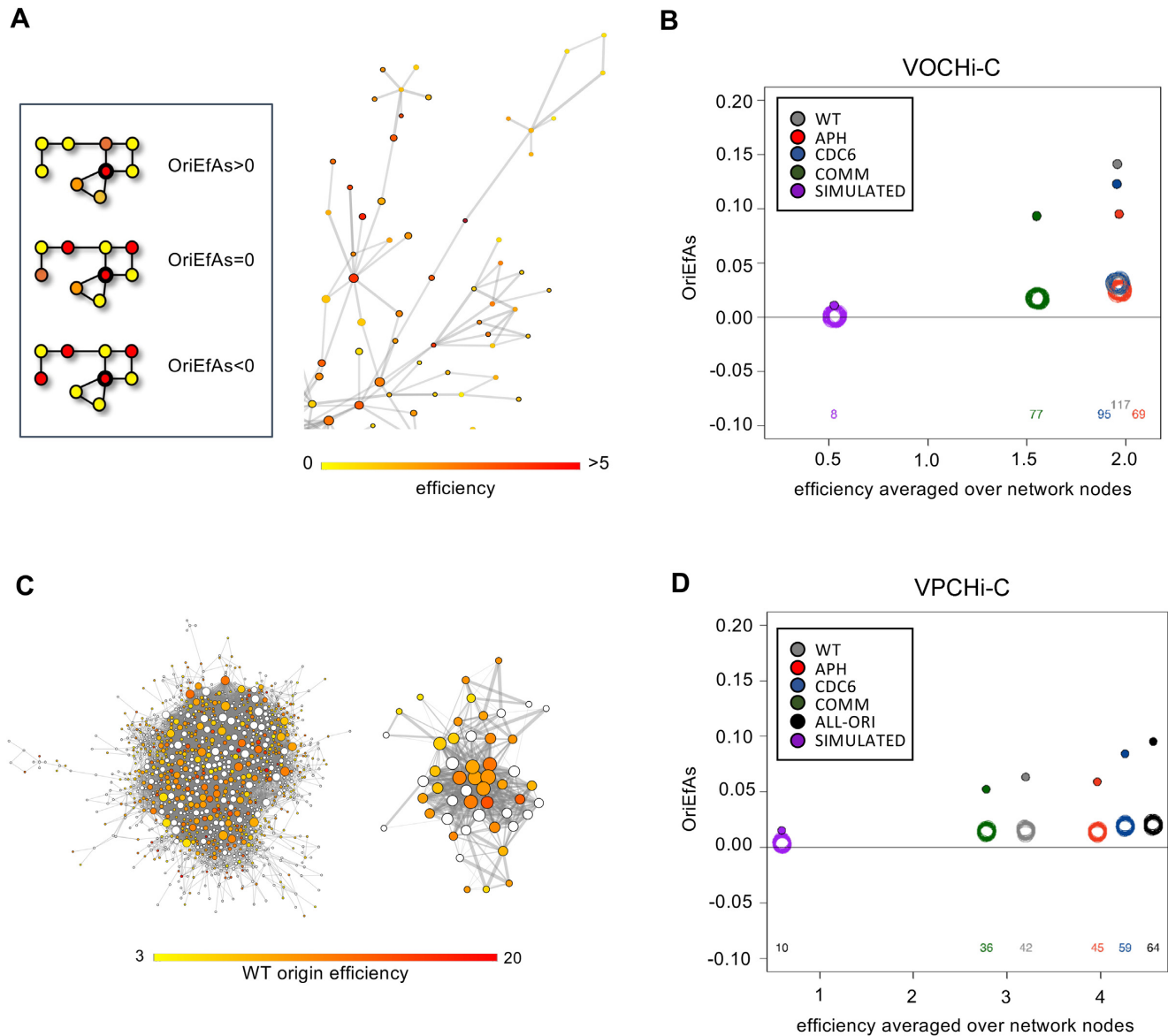
Figure 5. Origin efficiency is assortative in chromatin interaction networks. (**A**) Left, schematic of OriEfAs definition: positive OriEfAs indicates that origins of similar efficiency (colour) tend to interact with each other. Right, fragment of VOCHi-C network with nodes coloured by efficiency. Nodes with a black border denote COMM origins. Node size is proportional to its degree (number of connections). (**B**) OriEfAs in the VOCHi-C network. Filled small circles represent OriEfAs values for the indicated datasets. Colored large empty circles represent assortativity values for distance-preserving randomizations of the networks (Methods). Numbers at the bottom indicate the OriEfAs *Z*-scores values at each dataset, which estimate how different experimental values are from those expected at random. Purple circles represent assortativity values for simulated sets of origin positions and their respective randomizations, matching the distances to the closest TSS and expression values of the corresponding gene. For clarity, only one representative set of simulated origins is shown, out of 20 simulations. (**C**) Illustration of VPCHi-C network (chromosome 1) with node color representing origin efficiencies. White circles represent promoter nodes not containing an origin. Node size is proportional to node degree. The right blow-up section shows a subcluster of particularly interconnected nodes. (**D**) OriEfAs in VPCHi-C network. All symbols as in (**B**). One representative set of simulated origins (out of 20) is shown.

VPCHi-C nodes were defined as 5 kb regions surrounding the TSS and the interactions between them were extracted from the Hi-C matrix. VPCHi-C comprises 22 711 nodes and 841 869 edges, and it includes 40% of the origins in the WT dataset (Supplementary File 1). Node efficiencies were calculated as the average efficiency of origins within each fragment (Methods; Figure 5C). OriEfAs was significant in VPCHi-C for all origin datasets, including a combined set of all detected origins (ALL-ORI; OriEfAs range: 0.052–0.095, *Z*-scores 36–64; Figure 5D).

Common features of origins interacting across short- and long-range distances

The linear distance between interacting origins in VOCHi-C covered a broad range (<5 kb to 181 Mb; mean 25.9 Mb), with 94% of them spanning >1 Mb and 64% of them spanning >10 Mb. These predominantly long-range contacts imply that the majority of inter-origin interactions were detected across different topologically associated domains (TADs; green line in Supplementary Figure S8A). A

similar distribution was observed in the VPCHi-C network (Supplementary Figure S8B).

To analyze the properties of shorter-range origin connectivity networks, we used an alternative chromatin contact map generated by Promoter-Capture Hi-C (PCHi-C; (48)). PCHi-C identifies interactions spanning from 10 kb to >10 Mb between two promoters (P–P) or between a promoter and a non-promoter ('other end') region (P-O; Supp. File 1). In the subnetwork of PCHi-C chromatin fragments that contain at least one WT origin (PCHi-C-ORI), ~80% of the interactions were established within the same TAD, while the remaining 20% reflected inter-TAD links. The distribution of distances spanned by contacts in PCHi-C-ORI confirms the presence of a subset of long-range interactions (Supplementary Figure S8C; for a graphical representation, see Supplementary Figure S8D). As expected, origin efficiency was assortative in the PCHi-C P–P subnetwork (Supplementary Figure S8E).

We next analyzed a TAD-based contact network derived from VOCHi-C in which TAD efficiency was estimated as the average efficiency of the origins contained in them (Figure 6A). Remarkably, TAD efficiency was also assortative (TAD-OriEfAs = 0.25, with randomized values in the –0.03–0.1 range). TADs are numbered consecutively according to their linear position along chromosomes (47), so the highlighted network fragment (Figure 6A, middle panel) represents an example of non-adjacent, interacting TADs. We next checked whether origins located in different TADs that interact in 3D displayed similar timing of replication. Both VOCHi-C and PCHi-C-ORI networks displayed remarkable assortativity values of replication timing (RTAs = 0.57 and 0.46, respectively). The difference in RT between pairs of contacting regions in VOCHi-C was much smaller than the expected values for randomly chosen genomic regions, independently of the contacts being inter- or intra-TAD (Figure 6B; Supplementary Table S4). This result holds true when the random pairs of origins were taken exclusively from the A or B compartments, suggesting that the effect is not entirely explained by uniform RT values within each compartment. Origin efficiency was also assortative on compartment-specific subnetworks derived from VOCHi-C (Supplementary Figure S9). Taken together, these analyses suggest that interacting origins tend to fire with similar efficiencies across the cell population, and replicate synchronously regardless of whether they belong to the same TAD or separate TADs.

DISCUSSION

The activation of 'dormant' replication origins in response to stress is a central mechanism to maintain genomic stability in mammalian cells (32,33). Despite this fact, the characteristics and regulation of stress-responsive origins are still not completely understood. In this study, we have compared the activity of mESC replication origins in control conditions or under situations of mild RS induced by APH or CDC6 overexpression. As expected for a DNA polymerase inhibitor, APH slowed down fork progression and triggered a higher frequency of origin activation. In turn, the positive effect of CDC6 overexpression on origin firing is likely caused by the enhanced recruitment of origin-activating factors to the DNA (62). In this regard,

we note that CDC6 overexpression did neither induce origin re-firing nor DNA over-replication, in agreement with our previous observations that these effects require simultaneous CDC6 and CDT1 deregulation in mouse ESCs and tissues (35,53). The intersection of WT, APH and CDC6 origin datasets hinted at the existence of thousands of stress-responsive origins in the genome of mESCs. However, these responsive origins were also active in a smaller fraction of the cell population in the unchallenged S phase. Therefore, we propose that the main response to stress at the origin level does not consist in the selection of new initiation sites that were completely dormant, but in the modulation of the activity of pre-existing origins. It remains possible, however, that under conditions of stronger RS, a more extensive cellular response involving SIRT1 deacetylase may trigger the activation of otherwise silent origins, as recently shown in human somatic cancer cells (9).

Recent studies indicate that the selection of initiation sites has a marked stochastic component: potential origins are located at preferred genomic sites and each one has a given probability of activation. Stochastic origin activity was first described in unicellular yeasts (63,64) and has been supported by the technically groundbreaking optical mapping of initiation sites in mammalian cells, which suggests that even the most efficient origins are active in a fraction of the total cells (21). Our new data of the regulation of mESC initiation sites in response to RS also reflects a partially stochastic behavior. Common origins, i.e. those experimentally identified in every condition, are amongst the most efficient initiation sites and likely share characteristics with the human 'core origins' whose high activity is observed across different cell types (7). In turn, stress-responsive origins are initially used in a smaller fraction of cells but upon fork slowdown have extra time to be activated. We emphasize that the classical view of dormant origins as potential initiation sites that remain silent in the unchallenged S phase is still valid for individual cells, but they cannot be detected by SNS-Seq assays that compile data from millions of cells analyzed simultaneously (see model in Figure 7). Responsive origins could be found in every efficiency quantile, in line with the observation that even the most efficient origins in human cells can be further stimulated upon mild RS induced by hydroxyurea (12). This study and ours have reached consistent conclusions about stress-responsive origin activation, despite having used different cell types, RS-inducing drugs, origin mapping methods and efficiency calculation algorithms.

The most efficient origins in mESCs colocalized with chromatin features characteristic of active promoters, while the chromatin marks that correlate with the elongating form of RNAPII were amongst the least-enriched features. This may reflect evolutionary pressure to avoid or minimize transcription-replication conflicts. In this regard, it has been recently reported that RNA polymerase II may drive chromatin-bound MCM complexes to non-transcribed regions of the genome (65). Another hint of the pressure to separate both processes is that the fraction of mammalian origins functionally conserved in avian cells is 2–3-fold higher among those associated with CGI and TSS than in the rest (26).

Our study addresses how origin activation is influenced by the number of contacts established between chromatin
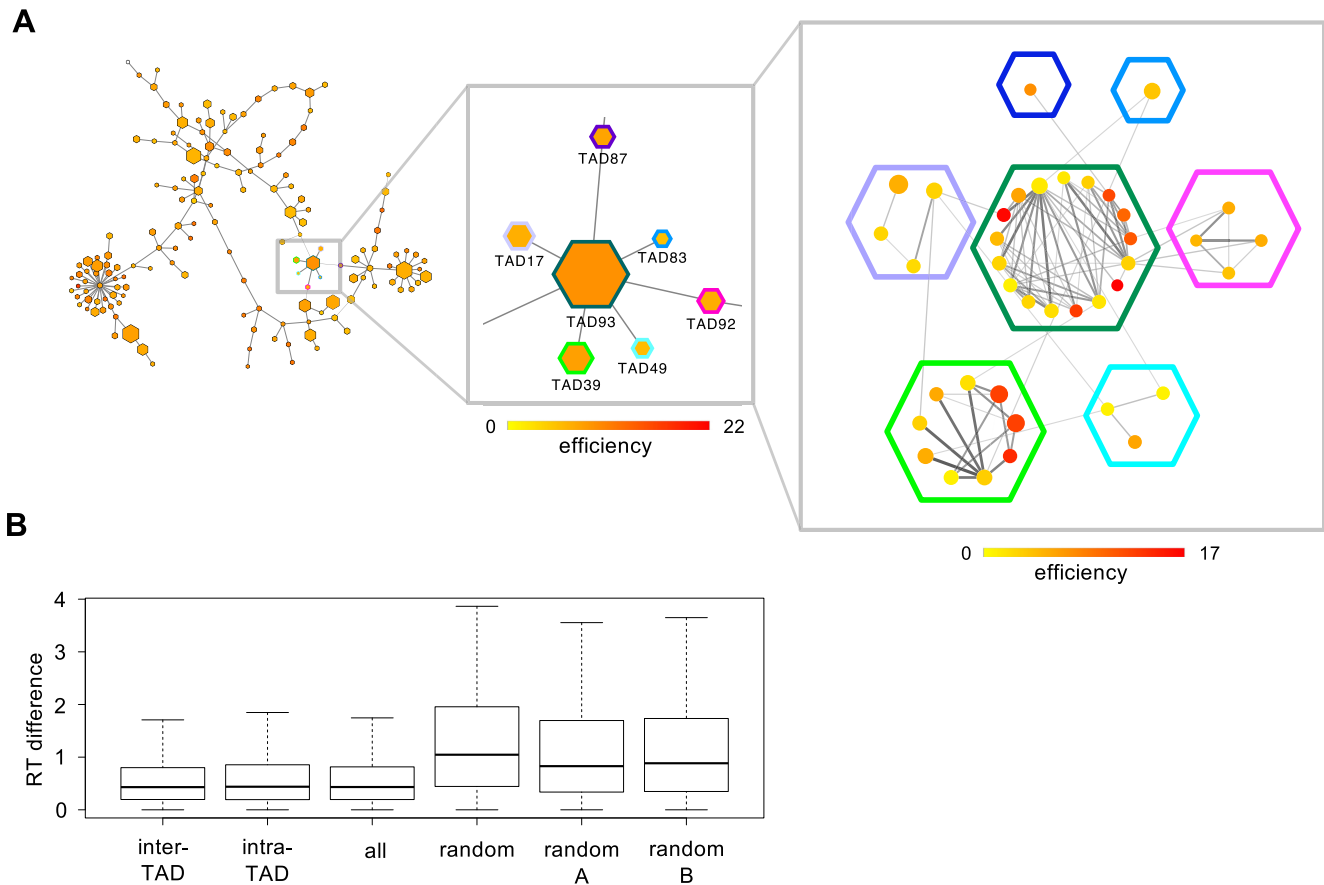
Figure 6. Interacting origins display similar efficiency and RT across a range of chromatin scales. (**A**) Left panel shows a subset of a TAD interaction network in which a yellow to red gradient in node color represents increasing origin efficiency (averaged within each TAD). Center panel shows a zoomed-in fragment of the network, in which the TAD identity of individual nodes is indicated. Right panel shows the origin interaction network within each TAD. In this case, node color represents origin efficiency. (**B**) Box plots show the RT difference between the indicated types of contacts within VOCHi-C. While all comparisons are significant ($P < 10^{-16}$, Wilcoxon), the differences in RT between inter-TAD, intra-TAD or all origins are much smaller than for randomized regions taken from the entire network or specifically from the A and B compartments (Supplementary Table S4).

fragments containing origins. We find that highly efficient origins tend to establish multiple connections between them, while lower efficiency origins establish fewer or no connections with other origin-containing chromatin regions. We are aware of an interesting precedent in which the application of graph theory to the analysis of Hi-C data led to the notion that 'master replication origins' are located at loci of maximal network centrality (66). We show here that the positive correlation between 3D connectivity and origin efficiency is also observed in subnetworks devoid of promoter nodes, suggesting that this correlation cannot solely be explained by the frequent overlap of origins with highly connected promoters. Furthermore, we report that origin efficiency and replication timing are assortative in three chromatin interaction networks (VOCHi-C, VPCHi-C and PCHi-C) that are different in terms of node number, fragment sizes, and number of interactions. Positive assortativity indicates that interacting origins tend to be activated with similar frequencies and replicated at the same time in S phase.

The abundance of origin-origin contacts supports the existence of DNA replication factories, a classic model for replication units in which clusters of origins are arranged

in physical proximity while inter-origin DNA is looped out (36,67). This architectural arrangement is facilitated by cohesin (68) and likely creates a favorable environment for the local concentration of origin-binding and origin-activating proteins. While replication factories are frequently represented as groups of origins that occupy adjacent positions along a chromosome and likely belong to a same chromosome structural unit, our analyses indicate that a large number of origin-origin contacts span very long distances, bridging together non-adjacent TADs. Although the correlations between initiation zones and TAD boundaries may not be identical in all cell types (69), a recent study using super-resolution optical imaging of replication foci has proposed that efficient origins are located at the periphery of TADs and could mediate inter-TAD contacts (70). Additional support for the strong association between 3D chromatin contacts and origin activation can be inferred from the delocalization of high-efficiency initiation zones when the cohesin-mediated loop extrusion mechanism is impaired (71).

Further analyses with chromatin network theory will shed additional light into the many interesting phenomena underlying the 3D organization of the DNA replication pro-
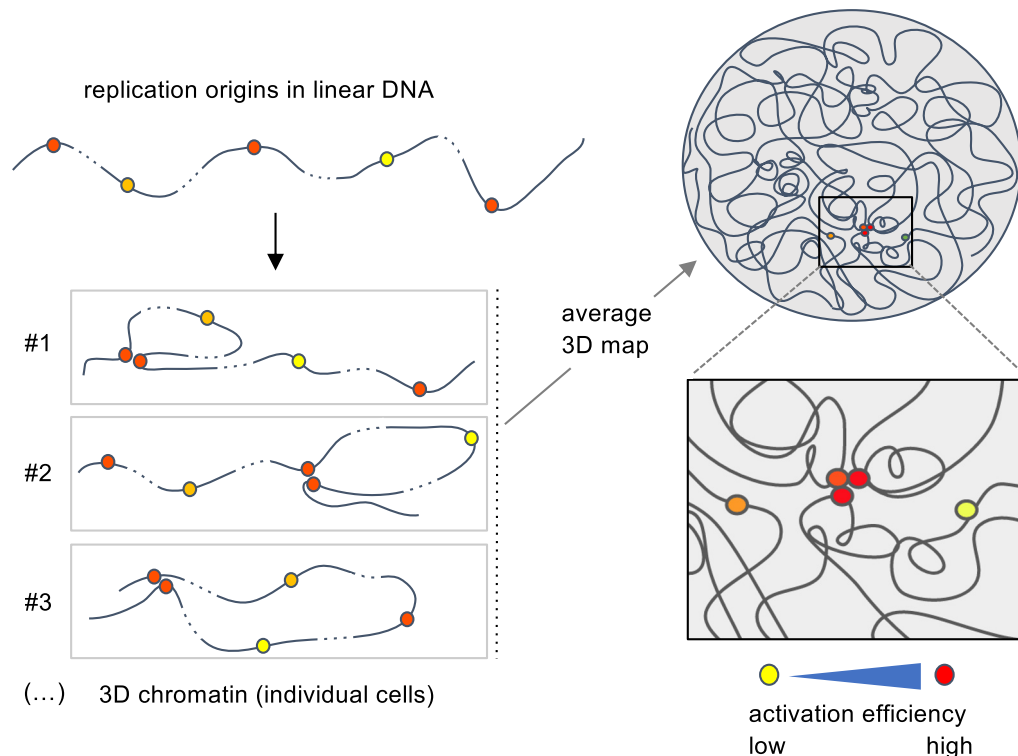
Figure 7. Schematic model representing the linear arrangement of a group of origins in a chromosomal fragment and how they may fold into 3D chromatin structure in different cells. On average, efficient origins tend to form connected hubs in the nucleus. See text for details.

cess, e.g. how the clustering of origins with promoters influences the RT program (72); how 3D genomic organization affects chromosome fragility (73); how RT is linked to the accumulation of mutations and copy number variations that determine the evolvability of specific genomic regions (74,75); and how origin activity and RT influence the frequency of oncogenic chromosomal translocations (76).

DATA AVAILABILITY

All sequence data generated in the study has been deposited in the Gene Expression Omnibus (GEO) database, with accession code GSE131699. Processed datasets used in the manuscript can be found at https://doi.org/10.6084/m9.figshare.c.2932049. Code and R notebooks can be found at https://github.com/VeraPancaldiLab/RepOri3D.The projections of origin efficiency datasets onto fragments of the PCHi-C network can be visualized at www.https://pancaldi.bsc.es/garden-net, where chromatin assortativity calculations can be performed.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. Aladjem,M.I. and Redon,C.E. (2017) Order from clutter: selective interactions at mammalian replication origins. *Nat. Rev. Genet.*, **18**, 101–116.
2. Prioleau,M.-N. and MacAlpine,D.M. (2016) DNA replication origins—where do we begin? *Genes Dev.*, **30**, 1683–1697.
3. Lombraña,R., Almeida,R., Revuelta,I., Madeira,S., Herranz,G., Saiz,N., Bastolla,U. and Gómez,M. (2013) High-resolution analysis of DNA synthesis start sites and nucleosome architecture at efficient mammalian replication origins. *EMBO J.*, **32**, 2631–2644.
4. Picard,F., Cadoret,J.C., Audit,B., Arneodo,A., Alberti,A., Battail,C., Duret,L. and Prioleau,M.N. (2014) The spatiotemporal program of DNA replication is associated with specific combinations of chromatin marks in human cells. *PLoS Genet.*, **10**, e1004282.
5. Cayrou,C., Ballester,B., Peiffer,I., Fenouil,R., Coulombe,P., Andrau,J.C., van Helden,J. and Méchali,M. (2015) The chromatin environment shapes DNA replication origin organization and defines origin classes. *Genome Res.*, **25**, 1873–1885.
6. Almeida,R., Fernández-Justel,J.M., Santa-María,C., Cadoret,J.C., Cano-Aroca,L., Lombraña,R., Herranz,G., Agresti,A. and Gómez,M. (2018) Chromatin conformation regulates the coordination between DNA replication and transcription. *Nat. Commun.*, **9**, 1590.
7. Akerman,I., Kasaai,B., Bazarova,A., Sang,P.B., Peiffer,I., Artufel,M., Derelle,R., Smith,G., Rodriguez-Martinez,M., Romano,M. et al. (2020) A predictable conserved DNA base composition signature defines human core DNA replication origins. *Nat. Commun.*, **11**, 4826.
8. Fu,H., Redon,C.E., Thakur,B.L., Utani,K., Sebastian,R., Jang,S.M., Gross,J.M., Mosavarpour,S., Marks,A.B., Zhuang,S.Z. et al. (2021) Dynamics of replication origin over-activation. *Nat. Commun.*, **12**, 3448.
9. Thakur,B.L., Baris,A.M., Fu,H., Redon,C.E., Pongor,L.S., Mosavarpour,S., Gross,J.M., Jang,S.M., Sebastian,R., Utani,K. et al. (2022) Convergence of SIRT1 and ATR signaling to modulate replication origin dormancy. *Nucleic Acids Res.*, **50**, 5111–5128.
10. Mesner,L.D., Valsakumar,V., Cieslik,M., Pickin,H., Hamlin,J.L. and Bekiranov,S. (2013) Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early- and late-firing origins. *Genome Res.*, **23**, 1774–1788.
11. Petryk,N., Kahli,M., D'Aubenton-Carafa,Y., Jaszczyszyn,Y., Shen,Y., Silvain,M., Thermes,C., Chen,C.L. and Hyrien,O. (2016) Replication landscape of the human genome. *Nat. Commun.*, **7**, 10208.
12. Chen,Y.H., Keegan,S., Kahli,M., Tonzi,P., Fenyö,D., Huang,T.T. and Smith,D.J. (2019) Transcription shapes DNA replication initiation and termination in human cells. *Nat. Struct. Mol. Biol.*, **26**, 67–77.
13. Langley,A.R., Gräf,S., Smith,J.C. and Krude,T. (2016) Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). *Nucleic Acids Res.*, **44**, 10230–10247.
14. Macheret,M. and Halazonetis,T.D. (2018) Intragenic origins due to short G1 phases underlie oncogene-induced DNA replication stress. *Nature*, **555**, 112–116.
15. Tubbs,A., Sridharan,S., van Wietmarschen,N., Maman,Y., Callen,E., Stanlie,A., Wu,W., Wu,X., Day,A., Wong,N. et al. (2018) Dual roles of Poly(da:dt) tracts in replication initiation and fork collapse. *Cell*, **174**, 1127–1142.
16. Guilbaud,G., Murat,P., Wilkes,H.S., Lerner,L.K., Sale,J.E. and Krude,T. (2022) Determination of human DNA replication origin position and efficiency reveals principles of initiation zone organisation. *Nucleic Acids Res.*, **50**, 7436–7450.
17. Dellino,G.I., Cittaro,D., Piccioni,R., Luzi,L., Banfi,S., Segalla,S., Cesaroni,M., Mendoza-Maldonado,R., Giacca,M. and Pelicci,P.G. (2013) Genome-wide mapping of human DNA-replication origins: levels of transcription at ORC1 sites regulate origin selection and replication timing. *Genome Res.*, **23**, 1–11.
18. Sugimoto,N., Maehara,K., Yoshida,K., Ohkawa,Y. and Fujita,M. (2018) Genome-wide analysis of the spatiotemporal regulation of firing and dormant replication origins in human cells. *Nucleic Acids Res.*, **46**, 6683–6696.
19. Miotto,B., Ji,Z. and Struhl,K. (2016) Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E4810–E4819.
20. Long,H., Zhang,L., Lv,M., Wen,Z., Zhang,W., Chen,X., Zhang,P., Li,T., Chang,L., Jin,C. et al. (2020) H2A.Z facilitates licensing and activation of early replication origins. *Nature*, **577**, 576–581.
21. Wang,W., Klein,K.N., Proesmans,K., Yang,H., Marchal,C., Zhu,X., Borrman,T., Hastie,A., Weng,Z., Bechhoefer,J. et al. (2021) Genome-wide mapping of human DNA replication by optical replication mapping supports a stochastic model of eukaryotic replication. *Mol. Cell*, **81**, 2975–2988.
22. Klein,K.N., Zhao,P.A., Lyu,X., Sasaki,T., Bartlett,D.A., Singh,A.M., Tasan,I., Zhang,M., Watts,L.P., Hiraga,S. et al. (2021) Replication timing maintains the global epigenetic state in human cells. *Science*, **372**, 371–378.
23. Sequeira-Mendes,J., Díaz-Uriarte,R., Apedaile,A., Huntley,D., Brockdorff,N. and Gómez,M. (2009) Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genet.*, **5**, e1000446.
24. Valton,A.-L., Hassan-Zadeh,V., Lema,I., Boggetto,N., Alberti,P., Saintome,C., Riou,J.-F. and Prioleau,M.-N. (2014) G4 motifs affect origin positioning and efficiency in two vertebrate replicators. *EMBO J.*, **33**, 732–746.
25. Prorok,P., Artufel,M., Aze,A., Coulombe,P., Peiffer,I., Lacroix,L., Guédin,A., Mergny,J.-L., Damaschke,J., Schepers,A. et al. (2019) Involvement of G-quadruplex regions in mammalian replication origin activity. *Nat. Commun.*, **10**, 3274.
26. Massip,F., Laurent,M., Brossas,C., Fernández-Justel,J.M., Gómez,M., Prioleau,M.N., Duret,L. and Picard,F. (2019) Evolution of replication origins in vertebrate genomes: rapid turnover despite selective constraints. *Nucleic Acids Res.*, **47**, 5114–5125.
27. Taylor,J.H. (1977) Increase in DNA replication sites in cells held at the beginning of s phase. *Chromosoma*, **62**, 291–300.
28. Besnard,E., Babled,A., Lapasset,L., Milhavet,O., Parrinello,H., Dantec,C., Marin,J.M. and Lemaitre,J.M. (2012) Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat. Struct. Mol. Biol.*, **19**, 837–844.
29. Comoglio,F., Schlumpf,T., Schmid,V., Rohs,R., Beisel,C. and Paro,R. (2015) High-Resolution profiling of drosophila replication start sites reveals a DNA shape and chromatin signature of metazoan origins. *Cell Rep.*, **11**, 821–834.
30. Muñoz,S. and Méndez,J. (2017) DNA replication stress: from molecular mechanisms to human disease. *Chromosoma*, **126**, 1–15.
31. Técher,H., Koundrioukoff,S., Nicolas,A. and Debatisse,M. (2017) The impact of replication stress on replication dynamics and DNA damage in vertebrate cells. *Nat. Rev. Genet.*, **18**, 535–550.
32. Ge,X.Q., Jackson,D.A. and Blow,J.J. (2007) Dormant origins licensed by excess mcm2–7 are required for human cells to survive replicative stress. *Genes Dev.*, **21**, 3331–3341.
33. Ibarra,A., Schwob,E. and Méndez,J. (2008) Excess MCM proteins protect human cells from replicative stress by licensing backup origins of replication. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 8956–8961.
34. Shima,N. and Pederson,K.D. (2017) Dormant origins as a built-in safeguard in eukaryotic DNA replication against genome instability and disease development. *DNA Repair (Amst.)*, **56**, 166–173.
35. Muñoz,S., Búa,S., Rodríguez-Acebes,S., Megías,D., Ortega,S., de Martino,A. and Méndez,J. (2017) In vivo DNA Re-replication elicits lethal tissue dysplasias. *Cell Rep.*, **19**, 928–938.
36. Jackson,D.A. and Pombo,A. (1998) 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.*, **140**, 1285–1295.
37. Gómez,M. and Antequera,F. (2008) Overreplication of short DNA regions during s phase in human cells. *Genes Dev.*, **22**, 375–385.

38. Foulk,M.S., Urban,J.M., Casella,C. and Gerbi,S.A. (2015) Characterizing and controlling intrinsic biases of lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-quadruplex motifs around a subset of human replication origins. *Genome Res.*, **125**, 725–735.

39. Cadoret,J.-C., Meisch,F., Hassan-Zadeh,V., Luyten,I., Guillet,C., Duret,L., Quesneville,H. and Prioleau,M.-N. (2008) Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 15837–15842.

40. Rubio-Camarillo,M., López-Fernández,H., Gómez-López,G., Carro,Á., Fernández,J.M., Torre,C.F., Fdez-Riverola,F. and Glez-Peña,D. (2017) RUbioSeq+: a multiplatform application that executes parallelized pipelines to analyse next-generation sequencing data. *Comput. Methods Programs Biomed.*, **138**, 73–81.

41. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

42. Feng,J., Liu,T., Qin,B., Zhang,Y. and Liu,X.S. (2012) Identifying chip-seq enrichment using MACS. *Nat. Protoc.*, **7**, 1728–1740.

43. Quinlan,A.R. (2014) BEDTools: the swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics*, **47**, 11.12.1–11.12.34.

44. Ye,T., Ravens,S., Krebs,A.R. and Tora,L. (2014) Interpreting and visualizing chip-seq data with the seqMINER software. *Stem Cell Transcriptional Netw.*, **1150**, 141–152.

45. Juan,D., Perner,J., Carrillo de Santa Pau,E., Marsili,S., Ochoa,D., Chung,H.-R., Vingron,M., Rico,D. and Valencia,A. (2016) Epigenomic co-localization and co-evolution reveal a key role for 5hmC as a communication hub in the chromatin network of ESCs. *Cell Rep.*, **14**, 1246–1257.

46. Hiratani,I., Ryba,T., Itoh,M., Rathjen,J., Kulik,M., Papp,B., Fussner,E., Bazett-Jones,D.P., Plath,K., Dalton,S. *et al.* (2010) Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome Res.*, **20**, 155–169.

47. Bonev,B., Mendelson Cohen,N., Szabo,Q., Fritsch,L., Papadopoulos,G.L., Lubling,Y., Xu,X., Lv,X., Hugnot,J.P., Tanay,A. *et al.* (2017) Multiscale 3D genome rewiring during mouse neural development. *Cell*, **171**, 557–572.

48. Schoenfelder,S., Furlan-Magaril,M., Mifsud,B., Tavares-Cadete,F., Sugar,R., Javierre,B.M., Nagano,T., Katsman,Y., Sakthidevi,M., Wingett,S.W. *et al.* (2015) The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.*, **25**, 582–597.

49. Cairns,J., Freire-Pritchett,P., Wingett,S.W., Várnai,C., Dimond,A., Plagnol,V., Zerbino,D., Schoenfelder,S., Javierre,B.-M., Osborne,C. *et al.* (2016) CHiCAGO: robust detection of DNA looping interactions in capture Hi-C data. *Genome Biol.*, **17**, 127.

50. Madrid-Mencía,M., Raineri,E., Cao,T.B.N. and Pancaldi,V. (2020) Using GARDEN-NET and ChAseR to explore human haematopoietic 3D chromatin interaction networks. *Nucleic Acids Res.*, **48**, 4066–4080.

51. Pancaldi,V., Carrillo-de-Santa-Pau,E., Javierre,B.M., Juan,D., Fraser,P., Spivakov,M., Valencia,A. and Rico,D. (2016) Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. *Genome Biol.*, **17**, 152.

52. Borlado,L.R. and Méndez,J. (2008) CDC6: from DNA replication to cell cycle checkpoints and oncogenesis. *Carcinogenesis*, **29**, 237–243.

53. Búa,S., Sotiropoulou,P., Sgarlata,C., Borlado,L.R., Eguren,M., Domínguez,O., Ortega,S., Malumbres,M., Blanpain,C. and Méndez,J. (2015) Deregulated expression of cdc6 in the skin facilitates papilloma formation and affects the hair growth cycle. *Cell Cycle*, **14**, 3897–3907.

54. Rodriguez-Acebes,S., Mourón,S. and Méndez,J. (2018) Uncoupling fork speed and origin activity to identify the primary cause of replicative stress phenotypes. *J. Biol. Chem.*, **293**, 12855–12861.

55. Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**, 817–825.

56. Lombraña,R., Álvarez,A., Fernández-Justel,J.M., Almeida,R., Poza-Carrión,C., Gomes,F., Calzada,A., Requena,J.M. and Gómez,M. (2016) Transcriptionally driven DNA replication program of the human parasite leishmania major. *Cell Rep.*, **16**, 1774–1786.

57. Marchal,C., Sima,J. and Gilbert,D.M. (2019) Control of DNA replication timing in the 3D genome. *Nat. Rev. Mol. Cell Biol.*, **20**, 721–737.

58. Sandhu,K.S., Li,G., Poh,H.M., Quek,Y.L.K., Sia,Y.Y., Peh,S.Q., Mulawadi,F.H., Lim,J., Sikic,M., Menghi,F. *et al.* (2012) Large-Scale functional organization of long-range chromatin interaction networks. *Cell Rep.*, **2**, 1207–1219.

59. Norton,H.K., Emerson,D.J., Huang,H., Kim,J., Titus,K.R., Gu,S., Bassett,D.S. and Phillips-Cremins,J.E. (2018) Detecting hierarchical genome folding with network modularity. *Nat. Methods*, **15**, 119–122.

60. Chovanec,P., Collier,A.J., Krueger,C., Várnai,C., Semprich,C.I., Schoenfelder,S., Corcoran,A.E. and Rugg-Gunn,P.J. (2021) Widespread reorganisation of pluripotent factor binding and gene regulatory interactions between human pluripotent states. *Nat. Commun.*, **12**, 2098.

61. Pancaldi,V. (2021) Chromatin network analyses: towards structure-function relationships in epigenomics. *Front. Bioinform.*, **1**, 742216.

62. Furstenthal,L., Kaiser,B.K., Swanson,C. and Jackson,P.K. (2001) Cyclin e uses cdc6 as a Chromatin-associated receptor required for DNA replication. *J. Cell Biol.*, **152**, 1267–1278.

63. Patel,P.K., Arcangioli,B., Baker,S.P., Bensimon,A. and Rhind,N. (2006) DNA replication origins fire stochastically in fission yeast. *Mol. Biol. Cell*, **17**, 308–316.

64. Czajkowsky,D.M., Liu,J., Hamlin,J.L. and Shao,Z. (2008) DNA combing reveals intrinsic temporal disorder in the replication of yeast chromosome vI. *J. Mol. Biol.*, **375**, 12–19.

65. Liu,Y., Ai,C., Gan,T., Wu,J., Jiang,Y., Liu,X., Lu,R., Gao,N., Li,Q., Ji,X. *et al.* (2021) Transcription shapes DNA replication initiation to preserve genome integrity. *Genome Biol.*, **22**, 176.

66. Boulos,R.E., Arneodo,A., Jensen,P. and Audit,B. (2013) Revealing long-range interconnected hubs in human chromatin interaction data using graph theory. *Phys. Rev. Lett.*, **111**, 118102.

67. Hozák,P., Hassan,A.B., Jackson,D.A. and Cook,P.R. (1993) Visualization of replication factories attached to a nucleoskeleton. *Cell*, **73**, 361–373.

68. Guillou,E., Ibarra,A., Coulon,V., Casado-Vela,J., Rico,D., Casal,I., Schwob,E., Losada,A. and Méndez,J. (2010) Cohesin organizes chromatin loops at DNA replication factories. *Genes Dev.*, **24**, 2812–2822.

69. Zhao,P.A., Sasaki,T. and Gilbert,D.M. (2020) High-resolution repli-seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome Biol.*, **21**, 76.

70. Li,Y., Xue,B., Zhang,M., Zhang,L., Hou,Y., Qin,Y., Long,H., Su,Q.P., Wang,Y., Guan,X. *et al.* (2021) Transcription-coupled structural dynamics of topologically associating domains regulate replication origin efficiency. *Genome Biol.*, **22**. 206.

71. Emerson,D.J., Zhao,P.A., Cook,A.L., Barnett,R.J., Klein,K.N., Saulebekova,D., Ge,C., Zhou,L., Simandi,Z., Minsk,M.K. *et al.* (2022) Cohesin-mediated loop anchors confine the locations of human replication origins. *Nature*, **606**, 812–819.

72. Brossas,C., Valton,A., Venev,S.v, Chilaka,S., Counillon,A., Laurent,M., Goncalves,C., Duriez,B., Picard,F., Dekker,J. *et al.* (2020) Clustering of strong replicators associated with active promoters is sufficient to establish an early-replicating domain. *EMBO J.*, **39**, e99520.

73. Sarni,D., Sasaki,T., Irony Tur-Sinai,M., Miron,K., Rivera-Mulia,J.C., Magnuson,B., Ljungman,M., Gilbert,D.M. and Kerem,B. (2020) 3D genome organization contributes to genome instability at fragile sites. *Nat. Commun.*, **11**, 3613.

74. Juan,D., Rico,D., Marques-Bonet,T., Fernández-Capetillo,Ó. and Valencia,A. (2014) Late-replicating CNVs as a source of new genes. *Biol Open*, **3**, 231–231.

75. Wu,X., Kabalane,H., Kahli,M., Petryk,N., Laperrousaz,B., Jaszczyszyn,Y., Drillon,G., Nicolini,F.-E., Perot,G., Robert,A. *et al.* (2018) Developmental and cancer-associated plasticity of DNA replication preferentially targets GC-poor, lowly expressed and late-replicating regions. *Nucleic Acids Res.*, **46**, 10157–10172.

76. Peycheva,M., Neumann,T., Malzl,D., Nazarova,M., Schoeberl,U.E. and Pavri,R. (2022) DNA replication timing directly regulates the frequency of oncogenic chromosomal translocations. *Science*, **377**, eabj5502.