Replication Stress Generates Multiple Distinct Classes of Copy Number Alterations.

Nadeem Shaikh†[1], Alice Mazzagatti†[1], Bjorn Bakker[2], Diana Carolina Johanna Spierings[2], René Wardenaar[4], Daniel Muliaditan[1], Eleni Maniati[1], Petter Larsson[1], Jun Wang[1], Floris Foijer[2], Sarah Elizabeth McClelland[1,4,*]

[1]Barts Cancer Institute, Queen Mary University of London, London EC1M 6BQ, UK. [2]European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, A. Deusinglaan 1, Groningen 9713 AV, the Netherlands.

†These authors contributed equally to this work.

*Corresponding author: s.mcclelland@qmul.ac.uk

Abstract

Replication stress can drive genetic instability and is associated with chromosomal instability in cancer and during reprogramming of stem cells, however the exact mechanisms connecting replication stress to chromosome aberrations are not known. Here we use single cell DNA sequencing to precisely map DNA copy number alterations (CNAs) after one cell division cycle under replication stress in two diploid cell types to megabase pair resolution. Unexpectedly, we find that replication stress generates several distinct classes of CNAs, and that specific genomic regions convert to particular classes of CNA. By combining analyses of single cell replication timing and bulk gene expression from both cell types, we provide a comprehensive picture of the features and mechanisms converting replication stress to DNA copy number alterations. The breakpoints of large CNAs tend to be associated with late DNA replication timing, low surrounding gene expression and proximity to large or giant genes. By contrast, small amplifications show contrasting associations and may represent a new class of replication stress induced CNAs. These findings provide a platform to further dissect molecular mechanisms involved in the replication stress response and to gain insights into how replication stress can drive chromosomal instability in cancer.

Introduction

Replication stress has been shown to drive chromosomal instability (CIN) in cancer[1,2] and induced pluripotent stem cells[3] and can be driven by oncogenes[4,5], low nucleoside concentrations[6,7] and difficult to replicate DNA sequences or structures[8,9]. High levels of replication stress can activate cell cycle checkpoints and lead to senescence that can form a barrier to tumour initiation[10,11]. Low levels however can bypass ATM and Chk1 activation[12] suggesting replication stress could act as a 'stealth' CIN mechanism in the presence of functional cellular checkpoints. Precisely how replication stress causes CIN remains unclear because the consequences on the genome of replication stress are only partially mapped, to date relying mainly on low-resolution cytogenetic analyses that detect breakpoint regions in metaphase spreads, termed common fragile sites (cFSs). Replication stress-induced DNA copy number alterations (CNAs) have also been detected using array comparative genomic hybridisation (aCGH) [13-15] which provides high resolution but is limited to the detection of those CNAs that survive clonal outgrowth. Replication stress induced CNAs mapped to date may therefore potentially represent only a minor fraction of the effects of replication stress on the genome.

To understand the precise and acute changes to the genome upon replication stress, we analysed CNAs induced after one cell cycle under replication stress caused by low dose aphidicolin, using single cell low pass whole genome sequencing, and in two diploid human cell types. This approach revealed multiple distinct classes of CNA, some of which were recurrent and clustered at characteristic genomic sites. A subset of large CNAs originated via chromosome breakage during mitosis, acentric lagging chromatin at anaphase, and acentric micronuclei in the following G1, explaining the aetiology of replication stress-induced chromosome segregation errors. Surprisingly, given the large number of previously-mapped cFSs in the genome, this subset was confined to only a few sites in RPE1 cells, resulting in the majority of acentric chromatin fragments and micronuclei comprising material from only three chromosomes. We performed single cell sequencing to determine DNA replication timing and bulk RNA sequencing to analyse gene expression in the same cell types, providing a complete picture of the factors involved in precipitating CNAs during replication stress from two different cell types. Large CNAs were generally associated with large or giant genes, low surrounding gene transcription levels and later replication timing. One highly recurrent fragile region of chromosome 7 in RPE1 cells appeared dependent on RPE1-specific gene transcription of a nearby giant gene, AUTS2. Small gains tended to be associated with higher gene transcription and early replication timing. With this novel approach therefore we have compiled a comprehensive catalogue of the CNAs induced acutely after replication stress, and begun to elucidate the mechanisms that link replication stress to chromosomal instability.

Results

Replication stress induced by low dose aphidicolin results in multiple classes of CNA that are associated with distinct genomic regions.

To identify CNAs induced by replication stress we treated two diploid, telomerase-immortalised human cell types; retinal pigment epithelial (RPE1-hTERT, hereafter RPE1) cells and foreskin fibroblasts (BJ-hTERT, hereafter BJ), with low doses of the DNA polymerase inhibitor aphidicolin for 24 hours. Induction of replication stress and genomic instability was verified by increases in replication stress-related DNA damage, namely γH2AX foci in prometaphase cells, segregation errors in anaphase and 53BP1 bodies in G1 cells as well as micronuclei in interphase cells (**Figure 1a-d; Figure S1a-l**). We then performed low pass single cell sequencing of G1 cells to detect whole and partial chromosome copy number changes[16-18] (**Figure 1e**) using EdU labelling to verify that cells in G1 had undergone DNA replication during that 24 h treatment period (**Figure S2a,b**). To gain greater depth of coverage we employed a modified library preparation protocol and sequenced at higher depth which generated ~ 0.04x coverage considering unique reads only (see **Methods** and **Supplemental Table 1**). This approach detected several classes of CNA, including a known sub-clonal trisomy of chromosome 12 and a clonal 10q amplification in RPE1 cells, and confirmed the XY karyotype of the BJ cells (**Figure 1f-g; Figure S2c**). In both cell types a set of recurrent focal amplifications were observed in both control and aphidicolin-treated cells. We reasoned that these likely represented existing structural variations present at clonal or sub-clonal frequencies since they were unique to each cell type. To verify that small amplifications were not simply general artefacts of the sequencing approach, we performed array comparative genomic hybridisation (aCGH) which verified the presence of a 2Mb focal amplification at chromosome 3p14.2 present in the majority of RPE1 cells (**Figure S2d**). We are thus able to

detect sub-chromosomal CNAs of 2 Mb with confidence using this approach. To visualise the landscape of replication stress-induced CNAs we removed clonal copy number variants, and other CNAs that occurred in or near centromeric regions that were likely mapping artefacts, and collated CNA events from 362 single RPE1 cells (**Figure 1h**). Aphidicolin treatment lead to an elevated rate of CNAs in both RPE1 and BJ cells with 50 % RPE1 cells and 57 % BJ cells exhibiting at least one aphidicolin-induced CNA (hereinafter 'aCNA') (**Figure 1i**). We noticed that aCNAs were either focal amplifications or deletions between 1-20 MB, or much larger amplifications or deletions (>20 MB) that often extended from the single breakpoint to the end of the chromosome arm ('large terminal aCNAs'; **Figure 1j,k**). Intriguingly, these distinct classes of aCNA tended to occur at different regions of the genome in both RPE1 and BJ cells (**Figure 1h; Figure S2c**). We hypothesised that different mechanisms may operate to drive these classes of lesion.

A small number of genomic sites are prone to replication-stress induced chromosome mis-segregation events and micronuclei containing a small subset of chromosomes.

Three regions, on chromosomes 1, 2 and 7, were recurrently prone to large terminal CNAs in RPE1 cells (**Figure 2a**). We wondered if these events could be connected with chromosome segregation errors observed following aphidicolin treatment[1] (**Figure 1a,b; Figure S1h,i**). Indeed, mis-segregating chromatin and micronuclei caused by aphidicolin lacked centromeres ('acentric') as judged by the lack of centromere-specific proteins or centromeric DNA sequences measured by immunofluorescence and fluorescence *In-Situ* hybridisation (FISH) respectively (**Figure 2b,c; Figure S3a-e**). Furthermore the majority (78%) of lagging chromatin fragments contained a focus of DNA damage at one, or both ends, as marked using antibodies to γH2AX (**Figure 2b,d**). We reasoned that if chromosome mis-segregation in RPE1 cells resulted from the large terminal aCNAs initiated at these recurrent breakpoints, then these chromosomes should be overrepresented in lagging chromosome fragments. Because anaphases are rare in asynchronous RPE1 cells, we analysed the identity of chromosomal material encapsulated within micronuclei that result from mis-segregating chromatin thus providing a record of previous chromosome segregation errors. FISH-based painting of specific chromosomes revealed that 18, 12 and 42 % of micronuclei in aphidicolin-treated RPE1 cells contained material from chromosomes 1, 2 or 7 respectively, whereas this enrichment was not observed in control cells (**Figure 2e,f**). This suggests that large terminal aCNAs are generated from breakage during mitosis, potentially as a result of condensation-induced rupture of under-replicated regions[19], followed by either mis-segregation into the wrong daughter cell (leading to gain of genetic material), or incorporation into micronuclei (leading to loss of genetic material). Since micronuclei are lost during preparation for single cell sequencing[17] this explains the bias towards losses of these large regions in single cell sequencing data (**Figure 2a**). These data are in agreement with a previous study that noted increased interphase aneuploidy of chromosomes harbouring common fragile sites in HCT116 and MRC5 cells[20].

Transcription of giant genes can underlie cell type-dependent susceptibility to large aCNAs.

We noticed that the region of chromosome 7q that was highly prone to large terminal aCNAs in RPE1 cells was less frequently altered in BJ cells (**Figure 3a**). Recurrent breakpoints in chromosomes 1 and 2 also followed this trend although to a lesser extent. These large aCNA-associated fragile regions were often obviously close to large

(>600 Kb) or giant (>1 Mb) genes (**Figure 3b**). Large genes have been previously associated with fragility under replication stress although the exact causes remain debated. It is known that common fragile site expression can vary between cell types and this is proposed to be a consequence of tissue-specific gene expression programs. To test whether gene expression differences at large and giant genes might underlie the cell-type dependent fragility on chromosome 7q, we performed RNA sequencing to determine the abundance of gene transcripts genome-wide in RPE1 and BJ cells treated with DMSO or aphidicolin. Total expression levels were broadly similar between the two cell lines (**Figure S4a**). In line with previous studies[13,21] aphidicolin treatment did not greatly impact gene expression globally, or of large or giant genes specifically (**Figure S4a**). AUTS1 and MAGI2 are giant genes close to the RPE1-specific aCNAs on chromosome 7q (**Figure 3b**). AUTS1 was expressed in RPE1, and not BJ cells (**Figure 3c**), suggesting that gene expression is a requirement for the fragility of this large-gene associated site, in line with a previous study[13]. However there was no clear effect of gene expression status at other large or giant genes within RPE1-specific recurrent large aCNA sites, and gene expression of many candidate genes was undetectable in both cell types (**Figure 3c**). We next analysed the genome-wide relationship between large aCNAs and proximity to large or giant genes for all classes of aCNA by plotting the distance to the nearest large or giant gene (**Figure 3d**). For large losses in both RPE1 and BJ cells, and large gains in BJ cells the mean distance to the nearest large gene was significantly lower compared to a set of randomly placed control regions. Additionally, large losses and gains in both cell lines exhibited populations of aCNAs that were very close to large or giant genes (**Figure 3d**). These observations suggest that proximity to large and giant genes appears to be linked to fragility but that steady state gene expression levels could not explain all differences in fragility between different cell types. To assess whether differences in replication timing of these regions could contribute to differences between RPE1 and BJ cells we performed single cell sequencing based replication timing analysis in a similar manner to a recent study[22], isolating cells from 4 separate S-phase fractions to increase the resolution of DNA timing (**Figure S4b-d**). Single cell replication timing profiles were similar to replication timing profiles derived from sequencing of bulk-populations (**Figure S4e)** and provided the opportunity to define the replication timing based on the proportion of cells that had replicated a given genomic region summed across the four quartiles of S-phase ('Replication timing factor', see **Methods**). This revealed that replication timing was similar between both cell types at all large genes analysed, including chromosome 7 genes AUTS2 and MAGI2 (**Figure 3e**), suggesting this was not a major determinant of fragility *per se*.

Focal amplifications occur in regions with higher transcriptional activity and earlier replication timing compared to other classes of aCNA.

The most frequent aCNAs observed in both cell lines were small amplifications (1-20 Mb) (**Figure 1j,k**). However the cause for these was not immediately apparent. Other studies have suggested a role for large genic transcription units in triggering hotspots of CNAs[13] but causes of small gains specifically are not clear. We therefore systematically searched for potential factors that could promote focal amplifications. We analysed the sum of gene expression within a 2 Mb window centred around breakpoints categorised into each CNA class. We then compared this to gene expression of 470 randomly placed breakpoints to serve as a control (**Figure 4a**). Most CNA categories (large losses, large gains, and small losses) were not significantly different to control regions but

tended towards lower gene expression. In contrast, focal amplifications exhibited a wide range of gene expression density with a mean significantly above that of control regions in both RPE1 and BJ cells (**Figure 4a**). We also plotted the replication timing of the genomic regions encompassing aCNA breakpoints. Small gains were again notably different to the other aCNA categories in that they were replicated significantly earlier than control regions whereas other aCNA classes tended to be replicated late (**Figure 4b**). Lastly, since the focal amplifications tended to be less obviously clustered in the same genomic regions, we wondered whether we had overlooked an association with proximity to large or giant genes, as had been observed for large terminal aCNAs (**Figure 3d**). Referring to this analysis we noticed that, in contrast to large gains and losses and focal losses, focal amplifications exhibited a similar distribution of proximity to large or giant genes to the randomly placed breakpoints (**Figure 3d**). We conclude from this that, as a class, small amplifications are associated with higher gene expression and earlier replication timing compared to large CNAs, and are not obviously associated with proximity to large or giant genes.

Discussion

Single cell sequencing of replication stress induced CNAs provides a complement to previous studies and can detect non-recurrent focal amplifications. Here we sought to comprehensively assess the impact that replication stress has on the genome. Previous studies tended to focus on a small number of recurrent fragile sites to understand molecular mechanisms operating to drive fragility under replication stress, potentially underlying some of the apparent conflicting views on the causes of fragility. A major advantage of our single cell sequencing approach is that we can follow not only large chromosomal breaks and gaps, but also the resulting copy number alterations induced at specific loci following replication stress at a resolution of at least 2 Mb. Replication stress-induced CNAs have been previously mapped to high genomic resolution using clonal outgrowth of cells where it was concluded that replication stress induced CNAs tend to occur at cFSs[13]. Our approach is additionally able to detect CNAs that might not be propagated in the population during clonal growth. We detected a large number of aphidicolin induced CNAs. We showed that large terminal aCNAs, often arising close to large or giant genes, convert into lagging chromosome material and micronuclei. These are likely to correspond to conventional cFSs. Additionally we detected a large number of small amplifications, many of which are unlikely to correspond to cFSs. First, many were not in recurrent positions and therefore would not be detected as common fragile sites (usually defined as representing at least 1% of cFSs). Second, the majority of aCNAs we detected were smaller than 20 Mb in size and therefore not detectable by cytological analyses. Moreover, small amplifications tended to occur in locations distinct from sites of large CNAs, suggesting they are caused by mechanisms distinct from those leading to previously mapped cFSs. A limitation of our study is that we cannot detect CNAs smaller than approximately 2 Mb with high confidence. We are also not currently able to sequence breakpoint junctions. Improving sequencing depth by using pre-amplification of single cell whole genomes would improve resolution to allow analysis of the sequences to reveal likely DNA repair mechanisms involved in generating aCNAs for example microhomology mediated mechanisms, or non-allelic homologous recombination.

What is special about recurrent sites of large terminal CNAs?

Large and giant genes have been proposed to cause fragility under replication stress due to a number of mechanisms including late[13,23-25] or delayed replication timing[21,26], lack of origin density[27], replication-transcription collisions during S-phase[5,13,28] and gene expression in G1 that removes licensing complexes and reduces origin density[26,29,30]. We showed that large terminal aCNAs were generally close to large or giant genes, and that in the case of the most highly recurrent site of aCNAs in RPE1 cells this was likely due to the RPE1-specific gene expression of the nearby giant gene AUTS2. However, later replication timing or steady state gene expression did not readily explain most cell type-dependent differences in fragility at recurrent large aCNAs (**Figure 3c-e**). We were also surprised by how few sites were prone to large terminal aCNAs, despite the presence of ~ 150 large or giant genes in the genome, many of which are expressed, and tend to replicate late[25]. There may be other factor(s) involved in predisposing these specific genomic sites to fragility. For example these genes may lie in regions that are replicated early or mid S-phase in normal cells but could be delayed in timing in the presence of aphidicolin[21,26] which would explain their fragility. The timing or rate of active gene transcription may also be an important factor that we were unable to detect with RNA sequencing. Performing alternative methods to measure active gene transcription is likely to shed further light on this. A second possibility is that variation in the population at the single cell level could account for fragility, since although these regions are considered recurrent they still only occur in 1-5% cells. Gene expression is known to vary widely between individual cells. Therefore single cell RNA sequencing may give an indication of whether, despite similarities at the bulk population level, variability in gene expression at the single cell level could explain why a subset of single cells are prone to specific aCNAs. Lastly, although we did not detect any obvious differences in replication timing at these large genes it is possible that delayed replication induced by replication stress could enhance fragility, as recently shown[21,26]. In our study we analysed replication timing from control-treated cells but it would be interesting to determine whether fragility of these regions correlates with aphidicolin-specific delays in replication timing. Our ability to characterise single cell-level differences would also provide interesting insights into the variability of such delays across individual cells.

What causes focal amplifications?

The majority of aCNAs detected in RPE1 and BJ cells were small amplifications. Although many of these were not in recurrent positions, these were infrequent in control-treated cells, and usually above 2 Mb in length, giving us confidence that the majority represent *bona fide* aphidicolin-induced CNAs. These small gains tended to be earlier replicated and within regions of higher gene expression that other CNA classes and the genome as a whole, but it is likely that additional factors are involved in promoting instability at these positions. Small gains may occur not as increased predisposition to fragility *per se*, but an increased propensity to be aberrantly repaired or dealt with under replication stress. For example underlying sequence features such as tandem duplications may promote aberrant recombinatorial repair. It is tempting to speculate that these aCNAs may represent a novel class of lesion induced by replication stress that has hitherto gone undetected. Determining the molecular mechanisms rendering these positions fragile, and the DNA repair and bypass mechanisms involved in their generation is therefore likely to provide new insights into the processes connecting replication stress to chromosomal instability.

Acknowledgments

We would like to thank Susana Godinho for kind gifts of reagents. We thank Daniel Muliaditan for generating randomly placed genomic regions. NS was funded by PCRF and Cancer Research UK. AM was funded by both PCRF and People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement n° 608765 . F.F. and B.B were funded by the Dutch Cancer Society grant 2012-RUG-5549.

Declaration of Interests

The authors declare no competing interests.

Author contributions

NS and AM designed and performed the majority of cell biological experiments and data analysis supervised by SEM. BB, RW and DJCS performed and analysed single cell sequencing data supervised by FF. DM created lists of random genome sites and input regarding bioinformatic analyses supervised by SEM. PL created python scripts to extract genomic features supervised by SEM. EM analysed RNA sequencing data supervised by JW. SEM conceived the study, designed experiments and wrote the manuscript with input from all authors.

Experimental procedures

Cell Culture

All cell lines were maintained at 37°C with 5% CO_2. hTERT-RPE-1 cells were cultured in DMEM Nutrient Mixture F12 Ham (Sigma); BJ cells in DMEM high glucose (Sigma). Media for both was supplemented with 10% FBS and 100 U Penicillin/Streptomycin. RPE1 and BJ cells were subjected to STR profiling to verify their identity using the cell line authentication service from Public Health England. Replication stress was induced using 0.2-0.4µM Aphidicolin (Sigma) for 24hrs.

Immunofluorescence (IF)

Cells grown on glass slides or coverslips were fixed with PTEMF (0.2% Triton X-100, 0.02 M PIPES (pH 6.8), 0.01 M EGTA, 1 mM $MgCl_2$, 4% formaldehyde). After blocking with 3% BSA, cells were incubated with primary antibodies according to suppliers' instructions: CREST (Antibodies Incorporated, 15-234-0001), γH2aX (Millipore, 05-636). Secondary antibodies used were goat anti-mouse AlexaFluor 488 (A11017, Invitrogen), goat anti-rabbit AF594, AF488 (A11012, A11008, Invitrogen), and goat anti-human AF647 (109-606-088-JIR, Stratech or A21445, Invitrogen). DNA was stained with DAPI (Roche) and coverslips mounted in Vectashield (Vector H-1000, Vector Laboratories). EdU incorporation and staining was achieved using the Click-It kit (Life Technologies), following manufacturer's instructions.

Fluorescence *In Situ* Hybridisation (FISH)

Cells were grown on glass slides, fixed in methanol/acetic acid, then put through an ethanol dehydration series. Pan-centromeric probe (Cambio) was denatured at 85°C for 10 minutes then applied to slides, which were then incubated in a humidified chamber overnight at 37C. The following day, slides were put through a series of washes (one 5 min wash at 37°C in 2xSSC, two 5 min washes at 37°C in 50% formamide/2xSSC, two 5 min washes at RT in 2xSSC). For chromosome painting, paint (Cytocell) was applied to the slide at 72°C for 2 min, then left overnight at 37°C in humidified chamber. The following day, slides were washed once with 0.4xSSC at 72°C for 2 min, then 2xSSC/0.05% Tween at RT for 30 s. After either staining methods, slides were then stained with DAPI, then coverslips with Vectashield were applied and sealed.

Microscopy

Images were acquired using an Olympus DeltaVision RT microscope (Applied Precision, LLC) equipped with a Coolsnap HQ camera. Three-dimensional image stacks were acquired in 0.2 µm steps, using Olympus 100× (1.4 numerical aperture), 60× or 40× UPlanSApo oil immersion objectives. Deconvolution of image stacks and quantitative measurements was performed with SoftWorx Explorer (Applied Precision, LLC). Analysis was performed using Softworx Explorer.

Single cell Sequencing

Samples from control and experimentally-induced aneuploid cells were sorted by FACS prior to next-generation sequencing library preparation and data analysis using AneuFinder as previously reported[16,31] except that the Strand Seq library preparation protocol[32] was used to create higher complexity libraries. Single nuclei were isolated and stained with 10 µg/mL propidium iodide and 10 µg/mL Hoechst. Single nuclei with low Hoechst/PI fluorescence (G1 population) were sorted into 96-well plates containing freezing buffer using a FACSJazz (BD Biosciences). Pre-amplification-free single-cell whole genome sequencing libraries were prepared using a Bravo Automated Liquid Handling Platform (Agilent Technologies, Santa Clara, CA, USA), followed by size-selection and extraction from a 2% E-gel EX (Invitrogen). Single-end 84 nt sequence reads were generated using the NextSeq 500 system (Illumina, San Diego, CA, USA) at 384 single-cell DNA libraries per flow cell. Demultiplexing based on library-specific barcodes and conversion to fastq format was done using bcl2fastq (v1.8.4, Illumina). Duplicate reads were called using BamUtil (v1.0.3). Demultiplexed reads were aligned to the GRCh38 reference genome using bowtie (v2.2.4) and only uniquely mapped reads (MAPQ>10) were used for further analysis. Copy number annotation was performed using AneuFinder (v1.4.0). Sequence reads are determined as non-overlapping bins with an average length of 1 Mb, a GC correction is applied, and binned sequences are analysed using a Hidden Markov model to determine the most likely copy number states. To negate the inherent sample variation introduced by sequencing single cells, a stringent quality control step was included that uses multivariate clustering to exclude libraries of insufficient quality. Chromosome copy number is plotted as a genome-wide state with clustering of cells based on the similarity of copy number profiles.

Generation of randomly placed control regions

Random genomic coordinates and 1 Mb intervals were generated using the bedtools random script from the BEDTools software[33]. 500 random coordinates were generated, 376 of which could be mapped to the genome during analyses of gene expression and replication timing.

RNA isolation, sequencing and analyses

Total RNA was extracted (RNeasy kit, Qiagen) from BJ and RPE1 cells, treated with either DMSO or aphidicolin, from three biological replicates. RNA quality was analysed on Tapestations, with RIN numbers were above 7.0. Library preparation was performed using the Lexogen 3' tagSeq kit. RNA sequencing was performed by Barts and the London Genome Centre on the Illumina NextSeq 500 platform, generating on average ~1 million single-end reads of 76 bp in length per sample. Raw reads were mapped to the human genome (hg38, Genome Reference Consortium GRCh38) using HISAT2[34]. Number of uniquely aligned reads ($q > 10$) to the exonic region of each gene were counted using HTSeq[35] based on the GenCode annotation release 29. Only genes that achieved at least one read count per million reads (cpm) in at least three samples were kept and a log2 transformed cpm expression matrix was subsequently generated. Differential expression analysis was performed using the 'limma' R package[36]. In order to analyse total gene expression levels at aCNAs in Figure 4 we generated a 2 Mb window of analysis centred around the putative breakpoint. The average expression of all genes in that window was summed. These values were then plotted in categories of small (<20 Mb) or large (>20 Mb) gain and loss CNAs. For large and giant gene expression analysis in Figure 3 we used the mean gene expression from three biological replicates.

Replication Timing

For replication timing both single and bulk G1 cells serving as controls were sorted. For experimental samples we sorted various S-phase populations based on DNA-content (late G1/early S; early S/mid S; mid S/late S; late S/early G2). Single-cell sequencing was performed and reads were processed as described above. The replication timing method as described by Takahashi[22] was then applied. Uniquely mappable sequence reads were binned into 1 Mb bins. Read counts per bin were then determined and converted into a read proportion across all reads. Bins with a read proportion lower than the 0.1 quantile were excluded. Quantiles were calculated separately for the autosomes and X-chromosome in male samples (the Y-chromosome was not included, neither were aneuploid chromosomes). To correct for variable mappability median centering was applied using the data from G1 cells as a reference. Correction factors were calculated by dividing the average read proportion by the proportion of each bin. Bin proportions were then multiplied by the correction factors to yield a corrected average proportion per bin. To determine whether a bin was replicated or not a quantile cut-off was applied per S-phase population to reflect the fraction of the genome expected to be replicated at that moment (i.e. 0.125, 0.375, 0.635, 0.875 for the population mentioned above). Bins with read proportion greater than the quantile cut-off set per S-phase population were designated as 'replicated' (blue in the plots). Conversely bins with proportions below the quantile cut-off were designated 'unreplicated' (yellow in the plots).

Replication timing Factor

For each bin, we determined whether the genome was replicated or not (1= replicated, 0= not replicated), in each of the single cells, at each of the four replication phases (early S phase, mid S phase, mid-late S phase and late S phase). We then summed the values across all replication phases, creating the replication factor value, where higher values represent earlier replication. We then categorised the range of replication factor values in four quartiles. Lowest quartile represents late replicated bins and highest quartile indicates bins that were replicated in early S phase, with mid-early and mid-late quartiles in between the two. To calculate large gene replication timing, we took the average replication factor for the two genomic bins that covered that large gene.

Array comparative genomic hybridisation (aCGH)

aCGH was performed by Aros AB (Denmark) using a Genome-Wide Human SNP Array 6.0 (Affymetrix) and data were analysed in the Chromosome Analysis Suite (CAS, Affymetrix). Data were transformed from global references obtained from signals in the CAS normalised reference library.

Statistical analysis

Unpaired *t*-test or one-way ANOVA Kruskal-Wallis test with post-hoc Dunn's correction were used to test for levels of significance using either Excel or Prism (GraphPad). Asterisks have been used to denote the significance value between experimental conditions adhering to the following nomenclature: $p<0.05$ (*); $p<0.005$ (**); $p<0.0005$ (***); $p<0.00005$ (****).

References

1 Burrell, R. A. *et al.* Replication stress links structural and numerical cancer chromosomal instability. *Nature* **494**, 492-496, doi:10.1038/nature11935 (2013).

2 Tamura, N. *et al.* Mechanisms of Chromosomal Instability in High-grade Serous Ovarian Carcinoma *BioRXiv* doi: http://dx.doi.org/10.1101/727537. (2019).

3 Ruiz, S. *et al.* Limiting replication stress during somatic cell reprogramming reduces genomic instability in induced pluripotent stem cells. *Nature communications* **6**, 8036, doi:10.1038/ncomms9036 (2015).

4 Miron, K., Golan-Lev, T., Dvir, R., Ben-David, E. & Kerem, B. Oncogenes create a unique landscape of fragile sites. *Nature communications* **6**, 7094, doi:10.1038/ncomms8094 (2015).

5 Jones, R. M. *et al.* Increased replication initiation and conflicts with transcription underlie Cyclin E-induced replication stress. *Oncogene* **32**, 3744-3753, doi:10.1038/onc.2012.387 (2013).

6 Bester, A. C. *et al.* Nucleotide deficiency promotes genomic instability in early stages of cancer development. *Cell* **145**, 435-446, doi:10.1016/j.cell.2011.03.044 (2011).

7 Foskolou, I. P. *et al.* Ribonucleotide Reductase Requires Subunit Switching in Hypoxia to Maintain DNA Replication. *Molecular cell* **66**, 206-220 e209, doi:10.1016/j.molcel.2017.03.005 (2017).

8 Kotsantis, P., Petermann, E. & Boulton, S. J. Mechanisms of Oncogene-Induced Replication Stress: Jigsaw Falling into Place. *Cancer discovery* **8**, 537-555, doi:10.1158/2159-8290.CD-17-1461 (2018).

9 Zeman, M. K. & Cimprich, K. A. Causes and consequences of replication stress. *Nature cell biology* **16**, 2-9, doi:10.1038/ncb2897 (2013).

10 Bartek, J., Bartkova, J. & Lukas, J. DNA damage signalling guards against activated oncogenes and tumour progression. *Oncogene* **26**, 7773-7779, doi:10.1038/sj.onc.1210881 (2007).

11 Bartkova, J. *et al.* Oncogene-induced senescence is part of the tumorigenesis barrier imposed by DNA damage checkpoints. *Nature* **444**, 633-637, doi:10.1038/nature05268 (2006).

12 Koundrioukoff, S. *et al.* Stepwise activation of the ATR signaling pathway upon increasing replication stress impacts fragile site integrity. *PLoS genetics* **9**, e1003643, doi:10.1371/journal.pgen.1003643 (2013).

13 Wilson, T. E. *et al.* Large transcription units unify copy number variants and common fragile sites arising under replication stress. *Genome research* **25**, 189-200, doi:10.1101/gr.177121.114 (2015).

14 Arlt, M. F., Ozdemir, A. C., Birkeland, S. R., Wilson, T. E. & Glover, T. W. Hydroxyurea induces de novo copy number variants in human cells. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 17360-17365, doi:10.1073/pnas.1109272108 (2011).
15 Arlt, M. F. *et al.* Replication Stress Induces Genome-wide Copy Number Changes in Human Cells that Resemble Polymorphic and Pathogenic Variants. *The American Journal of Human Genetics* **84**, 339-350, doi:10.1016/j.ajhg.2009.01.024 (2009).
16 Bakker, B. *et al.* Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome biology* **17**, 115, doi:10.1186/s13059-016-0971-7 (2016).
17 Worrall, J. T. *et al.* Non-random Mis-segregation of Human Chromosomes. *Cell reports* **23**, 3366-3380, doi:10.1016/j.celrep.2018.05.047 (2018).
18 Soto, M. *et al.* p53 Prohibits Propagation of Chromosome Segregation Errors that Produce Structural Aneuploidies. *Cell reports* **19**, 2423-2431, doi:10.1016/j.celrep.2017.05.055 (2017).
19 Lukas, C. *et al.* 53BP1 nuclear bodies form around DNA lesions generated by mitotic transmission of chromosomes under replication stress. *Nature cell biology* **13**, 243-253, doi:10.1038/ncb2201 (2011).
20 Minocherhomji, S. *et al.* Replication stress activates DNA repair synthesis in mitosis. *Nature* **528**, 286-290, doi:10.1038/nature16139 (2015).
21 Dan Sarni, T. S., Karin Miron, Michal Irony Tur-Sinai, Juan Carlos Rivera-Mulia, Brian Magnuson, Mats Lungman, David M. Gilbert and Batsheva Kerem. Replication Timing and Transcription Identifies a Novel Fragility Signature Under Replication Stress. *BioRXiv* http://dx.doi.org/10.1101/716951 (2019).
22 Takahashi, S. *et al.* Genome-wide stability of the DNA replication program in single mammalian cells. *Nature genetics* **51**, 529-540, doi:10.1038/s41588-019-0347-5 (2019).
23 Le Tallec, B. *et al.* Molecular profiling of common fragile sites in human fibroblasts. *Nature Structural & Molecular Biology* **18**, 1421-1423, doi:10.1038/nsmb.2155 (2011).
24 Blin, M. *et al.* Transcription-dependent regulation of replication dynamics modulates genome stability. *Nature structural & molecular biology* **26**, 58-66, doi:10.1038/s41594-018-0170-1 (2019).
25 Le Tallec, B. *et al.* Common fragile site profiling in epithelial and erythroid cells reveals that most recurrent cancer deletions lie in fragile sites hosting large genes. *Cell reports* **4**, 420-428, doi:10.1016/j.celrep.2013.07.003 (2013).
26 Olivier Brison, S. E.-H., Dana Azar, Stéphane Koundrioukoff, Mélanie Schmidt, Viola Naehse-Kumpf, Yan Jaszczyszyn, Anne-Marie Lachages, Bernard Dutrillaux, Claude Thermes, Michelle Debatisse and Chun-Long Chen. TRANSCRIPTION-MEDIATED ORGANIZATION OF THE REPLICATION INITIATION PROGRAM ACROSS LARGE GENES SETS UP COMMON FRAGILE SITES GENOME-WIDE. *BioRXiv* http://dx.doi.org/10.1101/714717 (2019).
27 Letessier, A. *et al.* Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. *Nature* **470**, 120-123, doi:10.1038/nature09745 (2011).
28 Helmrich, A., Ballarino, M. & Tora, L. Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Molecular cell* **44**, 966-977, doi:10.1016/j.molcel.2011.10.013 (2011).
29 Snyder, M., Sapolsky, R. J. & Davis, R. W. Transcription interferes with elements important for chromosome maintenance in Saccharomyces cerevisiae. *Molecular and cellular biology* **8**, 2184-2194, doi:10.1128/mcb.8.5.2184 (1988).
30 Looke, M. *et al.* Relicensing of transcriptionally inactivated replication origins in budding yeast. *The Journal of biological chemistry* **285**, 40004-40011, doi:10.1074/jbc.M110.148924 (2010).
31 van den Bos, H. *et al.* Single-cell whole genome sequencing reveals no evidence for common aneuploidy in normal and Alzheimer's disease neurons. *Genome biology* **17**, 116, doi:10.1186/s13059-016-0976-2 (2016).
32 Falconer, E. *et al.* DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nature methods* **9**, 1107-1112, doi:10.1038/nmeth.2206 (2012).
33 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
34 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nature methods* **12**, 357-360, doi:10.1038/nmeth.3317 (2015).
35 Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169, doi:10.1093/bioinformatics/btu638 (2015).
36 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43**, e47, doi:10.1093/nar/gkv007 (2015).

Figure 1

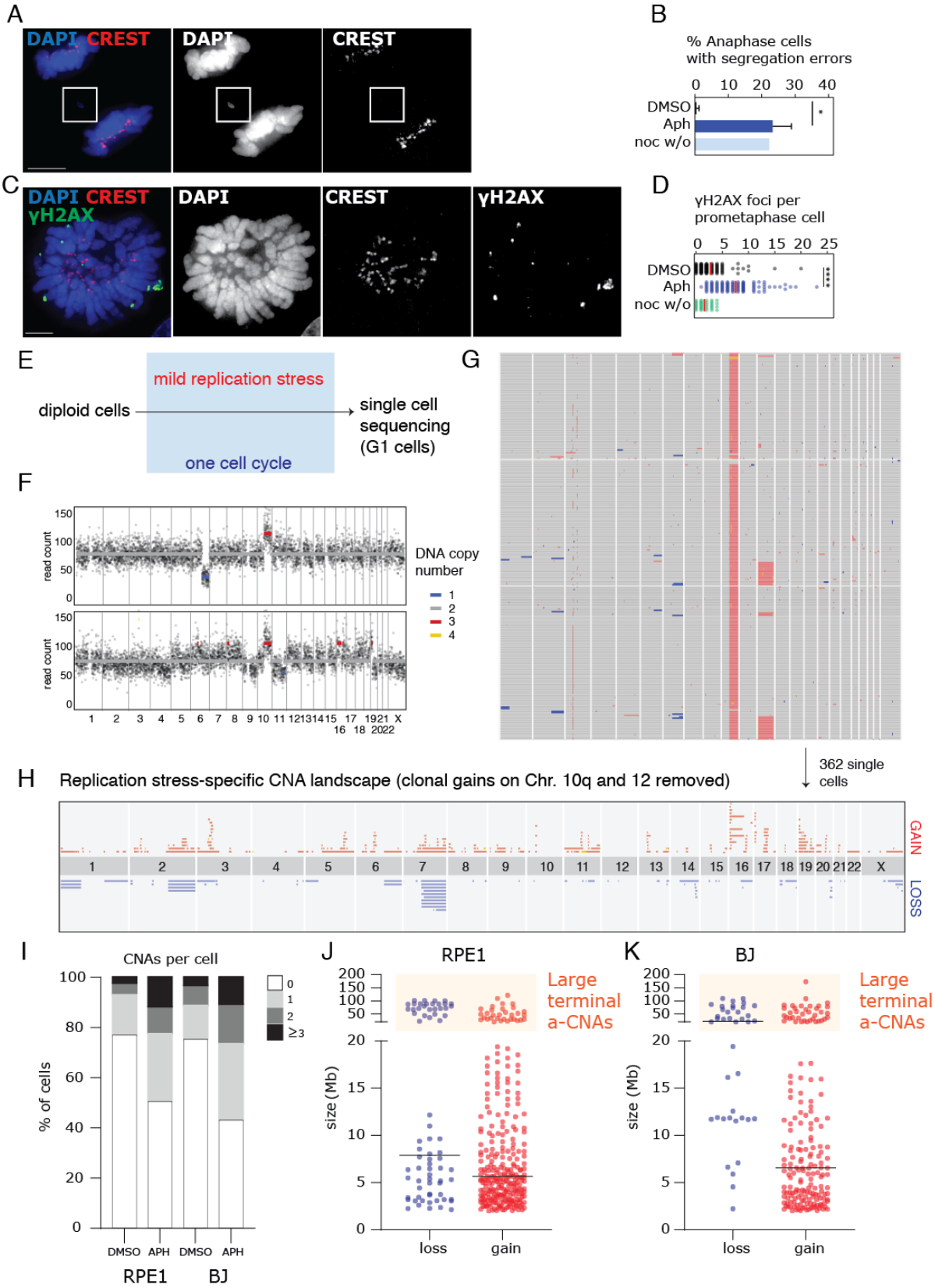

Figure 1: Replication stress induced by low dose aphidicolin results in multiple classes of CNA that are associated with distinct genomic regions. (A) Immunofluorescence image of RPE1 anaphase cell with lagging chromosome. CREST antibodies were used to stain for centromeric proteins. Scale bar in this and all subsequent microscopy

images represents 5 µM. (**B**) Segregation error rates in RPE1 anaphase cells after indicated treatments (summary of three experiments for 24 h DMSO and 24 h aphidicolin, one experiment for 4 h nocodazole with 1 h release; n=167, 150 and 36 cells, respectively). Red bar indicates mean and statistical test was an unpaired *t*-test. (**C**) Image of RPE1 prometaphase cell; DNA damage foci detected using γH2AX antibody. (**D**) Quantification of DNA damage in RPE1 prometaphase cells after indicated treatments (n=110, 105 or 37 cells, respectively). Statistical test was an unpaired *t*-test. (**E**) Schematic explaining experimental setup to generate Replication Stress CNAs. (**F**) Example of Single cell sequencing readout. Coloured bars represents average copy number for those reads. (**G**) Single Cell Sequencing data from 362 RPE1 cells treated with aphidicolin; blue represents copy number losses, red represents gains, grey indicates no copy number alteration. Amplification of 10q and recurrent aneuploidy of chromosome 12 are known aberrations in the RPE1 line. **(H)** Diagram summarising all RPE1 aCNAs taken from (G). (**I**) Quantification of frequency of CNAs in DMSO and aphidicolin treated RPE1 and BJ cells (n= 87, 362, 89 and 174 cells, respectively). (**J, K**) Distribution of aCNAs divided by size and by gains versus losses in RPE1 (J) and BJ (K) cells. Box highlights large terminal CNAs (>20 Mb) which span the chromosome arm up to and including the telomere.

Figure 2

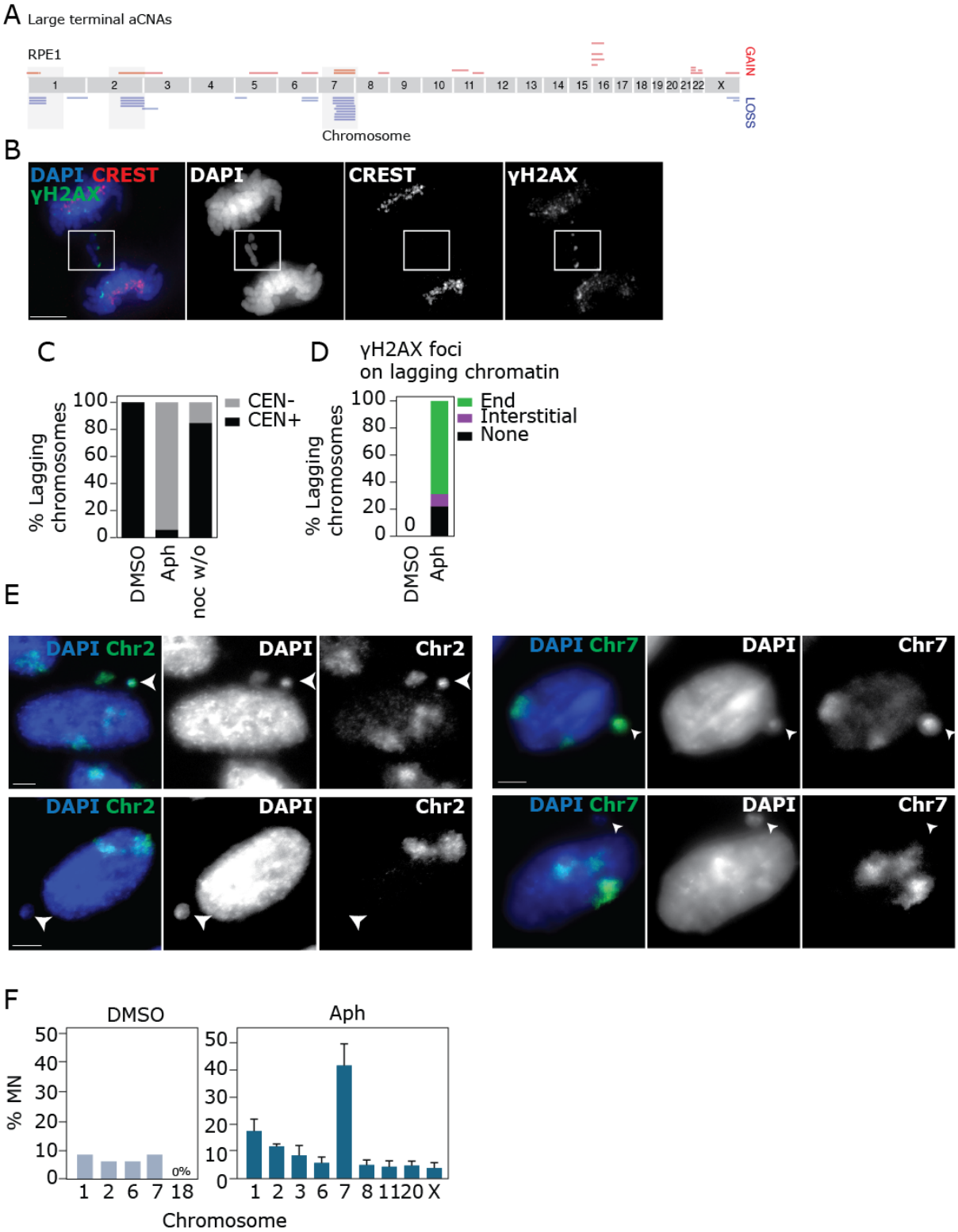

Figure 2: A small number of genomic sites are prone to replication-stress induced chromosome mis-segregation events and micronuclei containing a small subset of chromosomes.

(**A**) Schematic indicating location and relative sizes of large terminal aCNA gain and losses in RPE1 cells. (**B**) RPE1 anaphase cell with acentric lagging chromosome; centromeric proteins stained with CREST, DNA damage foci detected by γH2Ax staining. (**C**) Centromeric status of lagging chromosomes in RPE1 cells after indicated

treatments (24 h DMSO, 24 h aphidicolin, 4 h nocodazole 1 h release; n=2, 46 or 13 chromosomes, respectively). (**D**) Percentage of lagging chromosomes in RPE1 anaphase cells with DNA damage detectable on chromosome ends or within chromosome mass (n=35 chromosomes scored in aphidicolin). (**E**) Representative images of RPE1 cells stained with chromosome FISH paint to identify chromosomal identity of MN. (**F**) Quantification for positive staining of chromosome paints in RPE1 cells (DMSO one experiment, aphidicolin three experiments, 50-100 MN scored per chromosome per experiment).

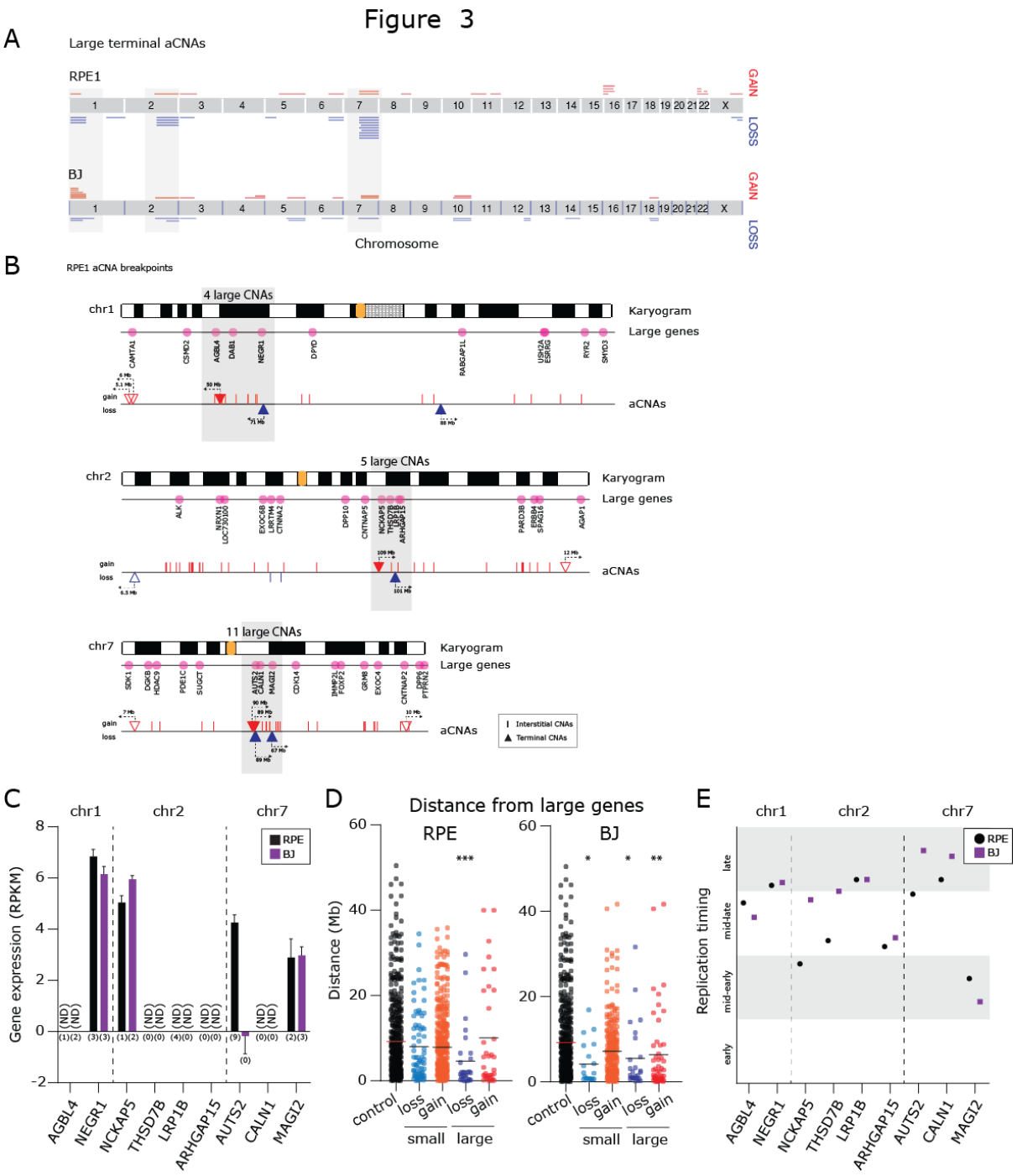

Figure 3 Transcription of giant genes can underlie cell type-dependent susceptibility to large aCNAs.
(**A**) Schematic indicating locations and frequency of large, terminal aCNAs in RPE1 and BJ cells. (**B**) Karyograms of three chromosomes with recurrent large, terminal aCNA in RPE1 cells, annotated with location of large genes. (**C**) Expression levels from both cell lines, of large genes nearest to recurrent large terminal aCNAs. (**D**) Distances from individual aCNAs to the nearest large or giant gene. Black or red lines indicate the mean. Control represents 470 randomly generated genomic co-ordinates. Statistical test compares all aCNA classes to control using a one way ANOVA Kruskal-Wallis test with post-hoc Dunn's correction. (**E**) Replication timing of large genes from (C) calculated as the mean of the replication timing factor of the. Genomic bin or bins (maximum two) containing the gene.

Figure 4

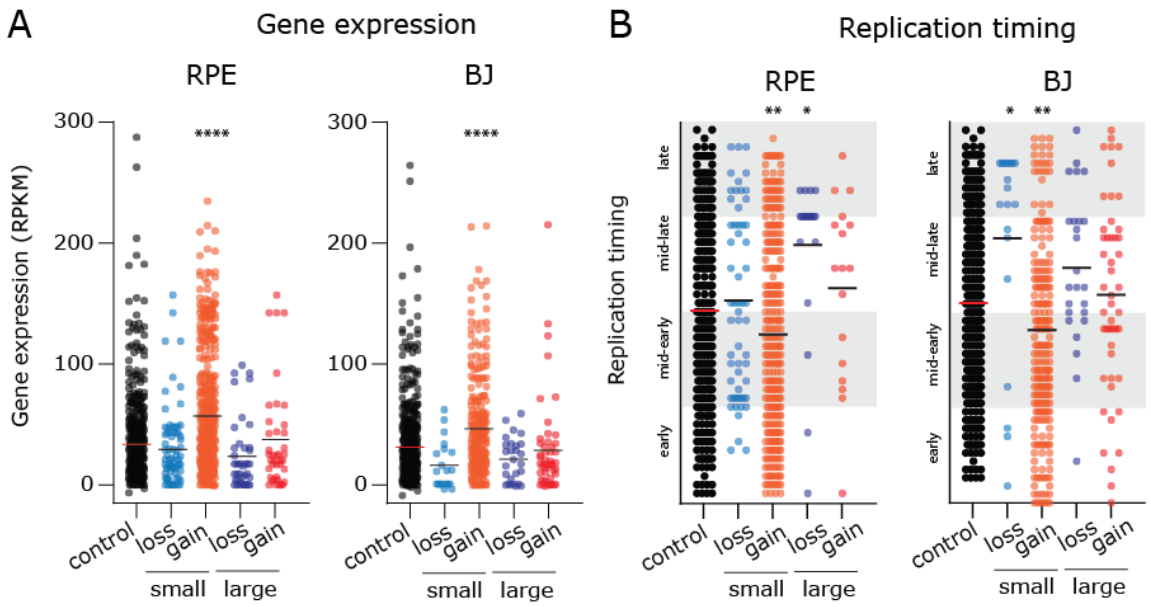

Figure 4 Focal amplifications occur in regions with higher transcriptional activity and earlier replication timing compared to other classes of aCNA.

(**A**) Summed gene expression within a 2 Mb window around each aCNA breakpoint, separated into CNA classes as indicated for RPE1 and BJ cells. Control data comes from 470 randomly generated 2 Mb windows, of which 376 were analysable for replication timing. (**B**) Replication timing factor was plotted for each aCNA in RPE1 and BJ cells. Statistical tests compare all aCNA classes to control using a one way ANOVA Kruskal-Wallis test with post-hoc Dunn's correction.

Figure S1

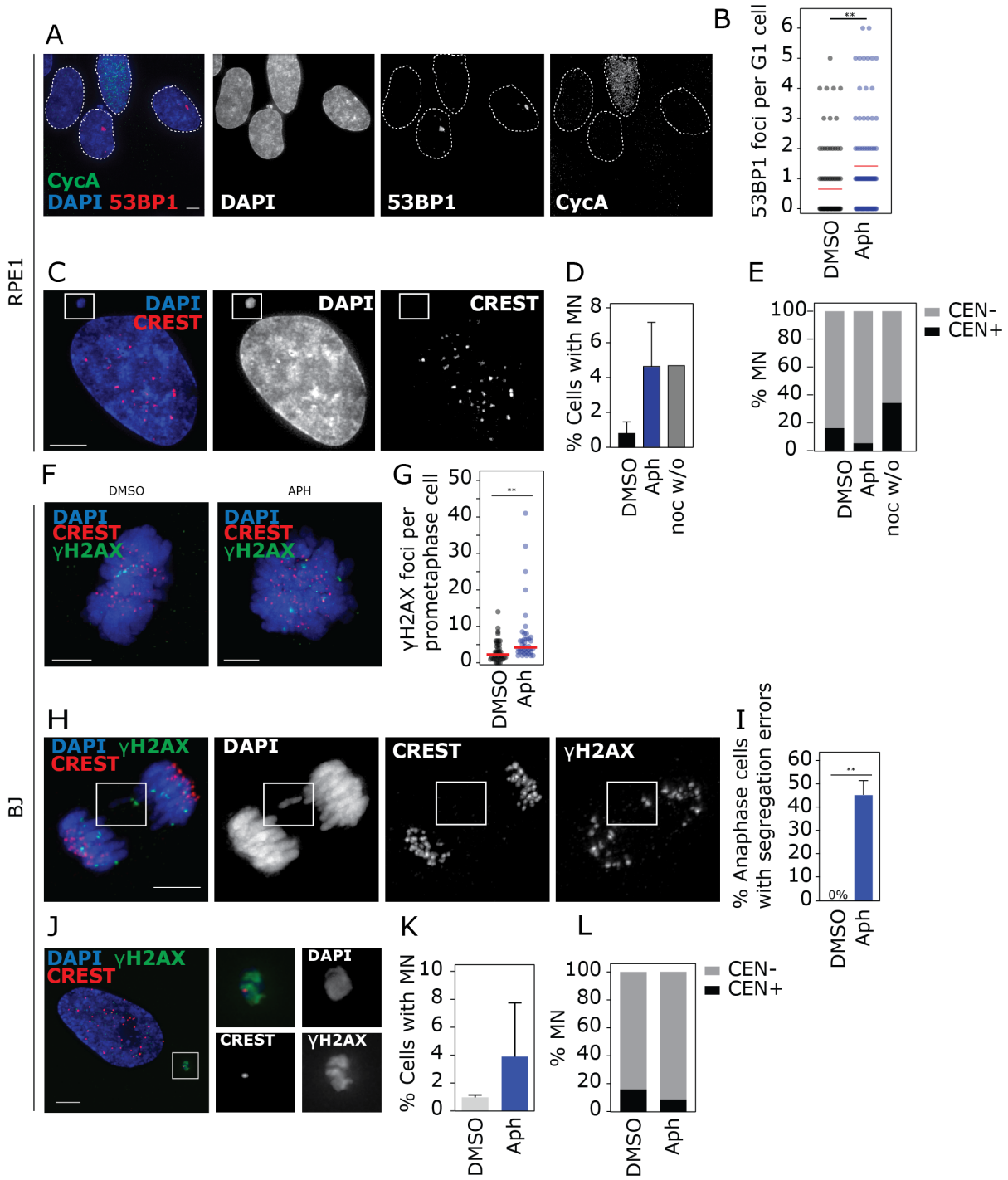

Figure S1: (relating to Figure 1): (A) Immunofluorescence image of RPE1 interphase cells, stained for 53BP1 bodies in G1 cells; Cyclin A staining used to identify S and G2 cells. **(B)** Quantification of 53BP1 bodies in G1 cells (n=106 and 107 cells). Red bar indicates mean. **(C)** Representative image of RPE1 cell with a micronucleus. CREST antibody was used to stain for centromeric proteins. **(D)** Quantification of MN rates in RPE1 cells after indicated treatments (three experiments for 24 h DMSO and 24 h aphidicolin, one for 4 h nocodazole and 16 h release; n=1029, 820 or 299 cells respectively). **(E)** Centromere status of MN in RPE1 cells after indicated treatments (n=94, 114 or 38 micronuclei, respectively), as determined by CREST staining. **(F)** Immunofluorescence images of BJ prometaphase cells. **(G)** Quantification of γH2Ax DNA damage foci in BJ cells after indicated treatments (n=66 and 73 cells). **(H)** Representative images of BJ anaphase cells with segregation errors. **(I)** Quantification of segregation errors in BJ anaphase cells. **(J)** Image of BJ interphase cell with micronucleus **(K)** MN rates in BJ cells after indicated treatments (n=743 and 600 cells). **(L)** Centromere status of BJ micronuclei. Statistical tests are unpaired *t*-tests.

Figure S2

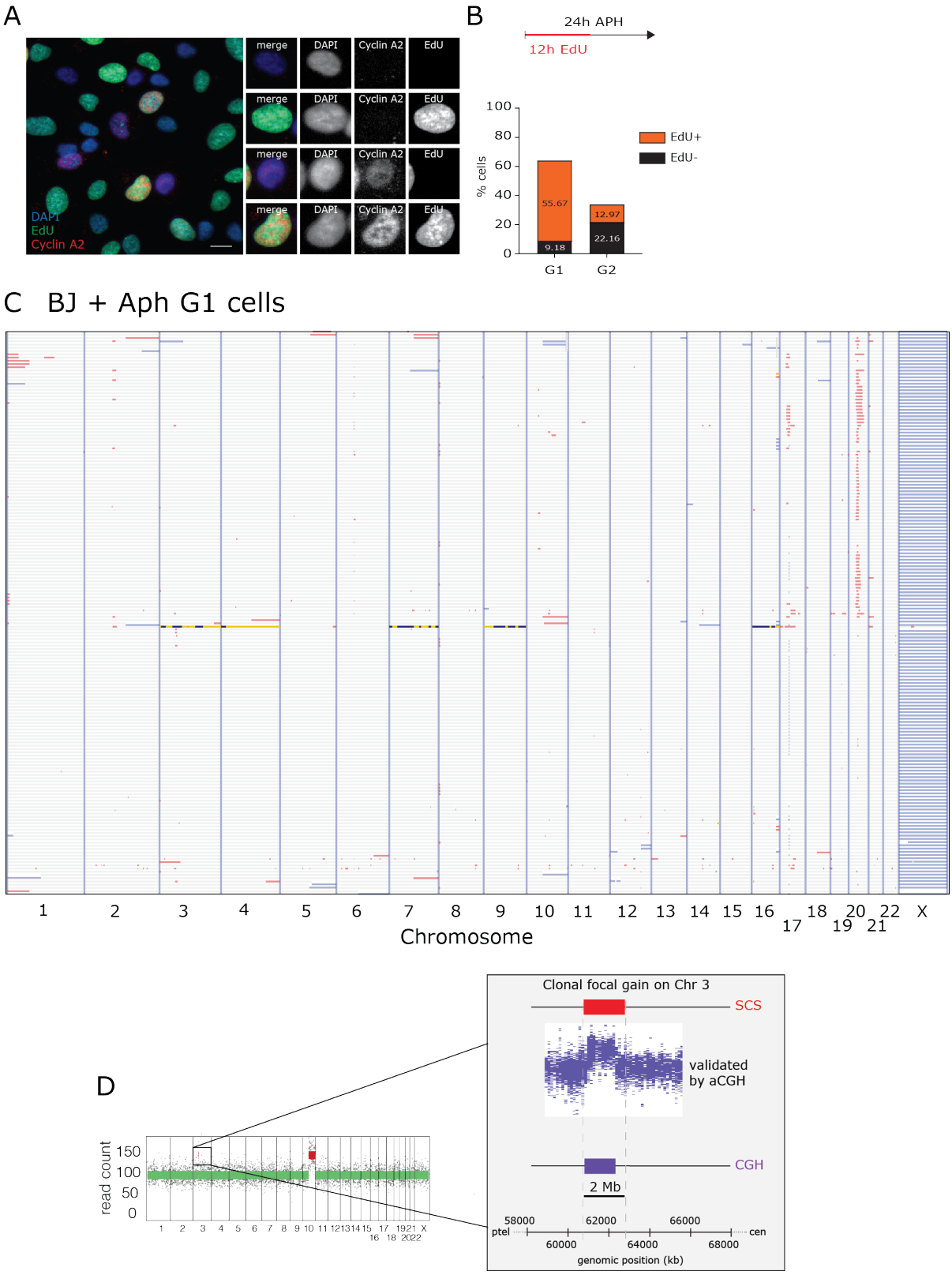

Figure S2: (relating to Figure 1) (**A**) Images of RPE1 cells stained for Cyclin A and EdU incorporation.(**B**) RPE1 cells were treated with DMSO or aphidicolin for 24 h, including a 12 h pulse of EdU as indicated. EdU incorporation into Cyclin A positive (S/G2) or negative cells (G1) was then quantitated from immunofluorescence microscopy images (n=185 and 235 cells). (**C**) Single cell sequencing data from 174 BJ cells after 24 h treatment with

aphidicolin. (**D**) Schematic indicating array CGH verification of clonal focal amplification on chromosome 3 in RPE1 cells, with size of detected amplification and location relative to centromere and p-telomere.

Figure S3

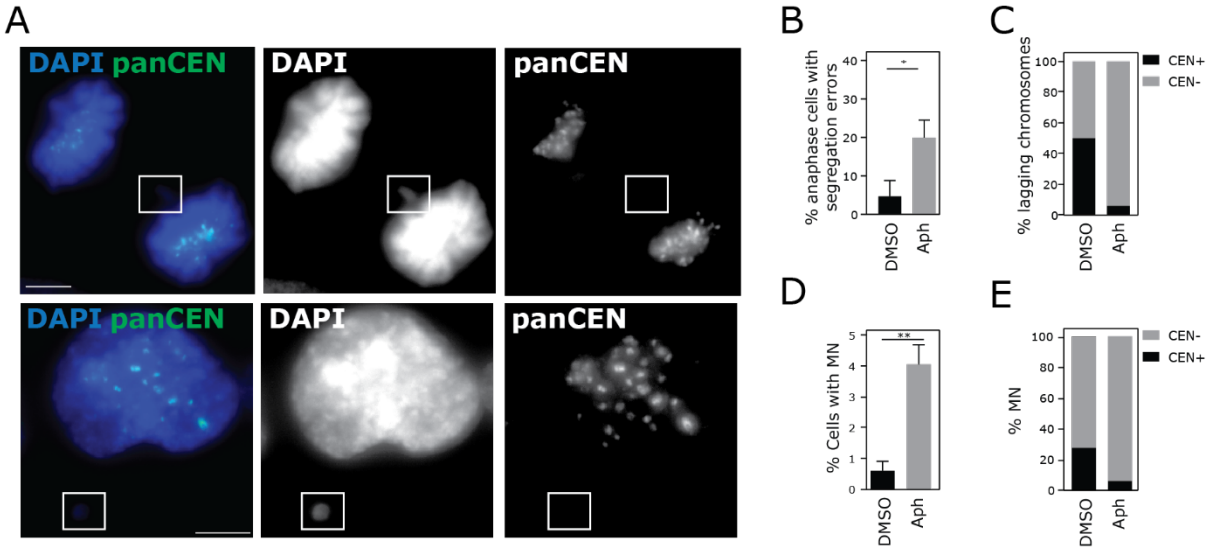

Figure S3: (relating to Figure 2): (**A**) Examples of RPE1 anaphase cells with lagging chromosomes, or interphase cells with MN, stained with pan-centromeric FISH probes. (**B**) Segregation error rates in RPE1 cells stained with FISH probes (n=106-109 cells). (**C**) Centromere status of lagging chromosomes in RPE1 cells as determined by pan-centromeric FISH staining (n=7 or 29 lagging chromosomes). (**D**) MN rates in RPE1 cells (n=725 or 674 cells). (**E**) Centromeric status of MN scored in RPE1 cells after pan-centromeric FISH (n=104-119 MN scored). Statistical tests are unpaired *t*-tests.

Figure S4

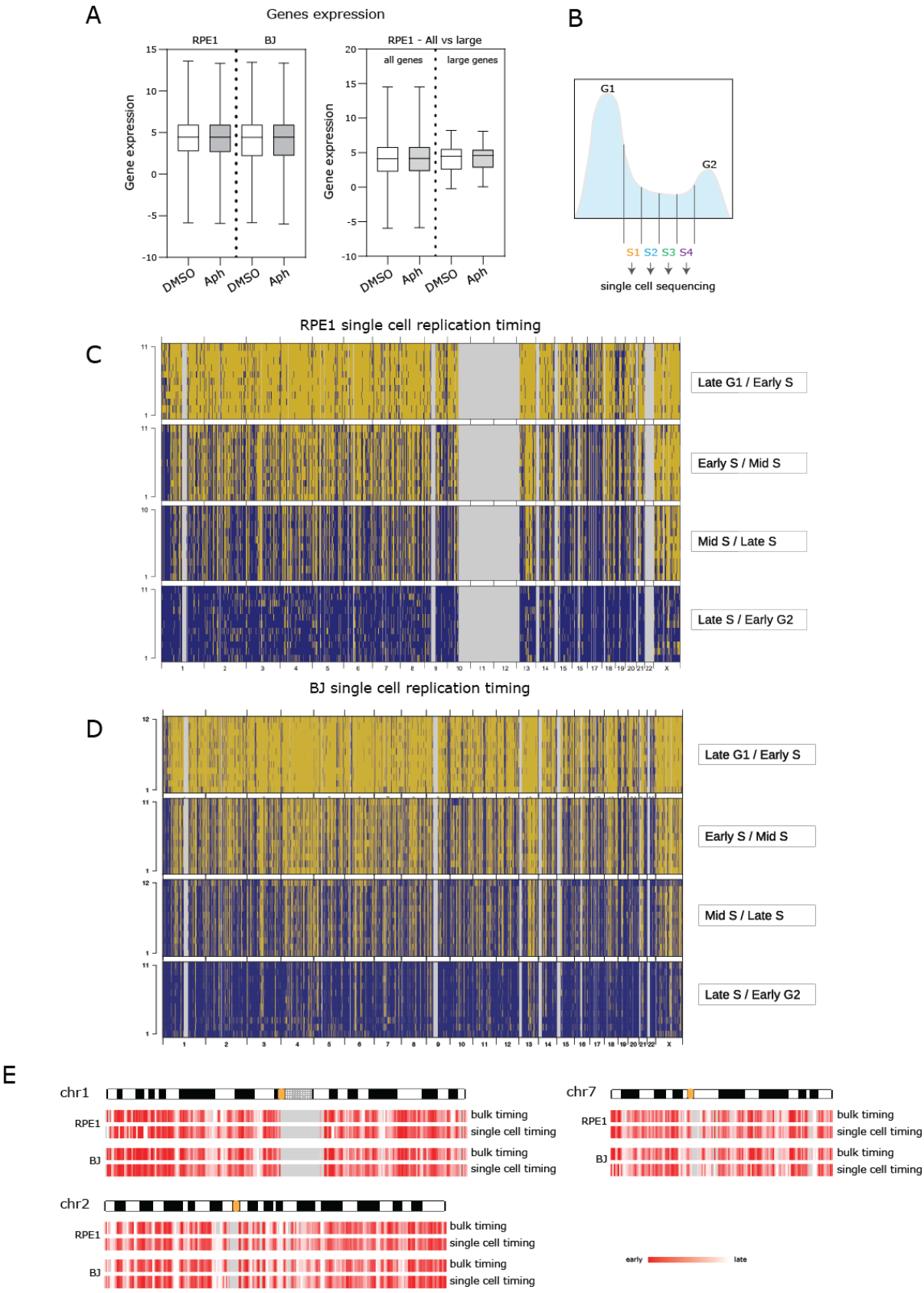

Figure S4 (related to Figure 3): (A) Left side: Total gene expression in DMSO vs aphidicolin for RPE1 and BJ cells; right side: expression in RPE1 cells treated with DMSO vs aphidicolin for all genes vs only large genes (>600 kb).

(**B**) Schematic indicating S-phase fractions isolated using FACS for single cell replication timing. (**C,D**) Single cell replication timing analyses for RPE1 (C) and BJ (D) cells. (**E**) Comparison between single cell and bulk replication timing for RPE1 and BJ cells for three selected chromosomes.