1280–1294 Nucleic Acids Research, 2018, Vol. 46, No. 3 *Published online 14 December 2017*
doi: 10.1093/nar/gkx1260

FANCD2 binding identifies conserved fragile sites at large transcribed genes in avian cells

Constanze Pentzold[1], Shiraz Ali Shah[1], Niels Richard Hansen[2], Benoît Le Tallec[3], Andaine Seguin-Orlando[4,5], Michelle Debatisse[6], Michael Lisby[1,7] and Vibe H. Oestergaard[1,*]

[1]Department of Biology; University of Copenhagen; Copenhagen N 2200, Denmark, [2]Department of Mathematical Sciences, University of Copenhagen, Copenhagen 2100, Denmark, [3]Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS-UMR8197 – Inserm U1024, Paris F-75005, France, [4]Center for GeoGenetics, Natural History Museum of Denmark; University of Copenhagen; Copenhagen 1350, Denmark, [5]Danish National High-throughput DNA Sequencing Centre, University of Copenhagen, Øster Farimagsgade 2D, Copenhagen K 1353, Denmark, [6]Institut Gustave Roussy, UMR 8200, 94800 Villejuif, France and [7]Center for Chromosome Stability, Department of Cellular and Molecular Medicine, University of Copenhagen, Blegdamsvej 3b, DK-2200 Copenhagen N, Denmark

Received September 27, 2017; Revised December 01, 2017; Editorial Decision December 05, 2017; Accepted December 05, 2017

ABSTRACT

Common Chromosomal Fragile Sites (CFSs) are specific genomic regions prone to form breaks on metaphase chromosomes in response to replication stress. Moreover, CFSs are mutational hotspots in cancer genomes, showing that the mutational mechanisms that operate at CFSs are highly active in cancer cells. Orthologs of human CFSs are found in a number of other mammals, but the extent of CFS conservation beyond the mammalian lineage is unclear. Characterization of CFSs from distantly related organisms can provide new insight into the biology underlying CFSs. Here, we have mapped CFSs in an avian cell line. We find that, overall the most significant CFSs coincide with extremely large conserved genes, from which very long transcripts are produced. However, no significant correlation between any sequence characteristics and CFSs is found. Moreover, we identified putative early replicating fragile sites (ERFSs), which is a distinct class of fragile sites and we developed a fluctuation analysis revealing high mutation rates at the CFS gene *PARK2*, with deletions as the most prevalent mutation. Finally, we show that avian homologs of the human CFS genes despite their fragility have resisted the general intron size reduction observed in birds suggesting that CFSs have a conserved biological function.

INTRODUCTION

During mitosis, a copy of the genome is passed on to each of the two daughter cells. Mitosis constitutes a short window of the entire cell cycle, but is of immense importance for genome integrity. In response to severe replication stress, the G2/M checkpoint will normally arrest cells in G2 phase but under conditions of mild replication stress, certain regions in the genome escape checkpoint detection and enter mitosis in an underreplicated state (1–5). This leads to the formation of microscopically visible breaks and gaps on metaphase chromosomes (6). Genomic regions prone to form gaps and breaks in response to replication stress are named chromosomal fragile sites (6). Interestingly, chromosomal fragile sites are hotspots for large deletions and rearrangements in cancer genomes (7–10). The genome of healthy individuals contains two classes of chromosomal fragile sites; (i) common fragile sites (CFSs) and (ii) early replicating fragile sites (ERFSs) (8,11). Underreplicated regions that persist in mitosis pose a problem to disjunction of sister chromatids and accordingly CFSs can remain interlinked by ultrafine DNA bridges (UFBs) or chromatin bridges during anaphase (3,12,13).

Since CFSs were first discovered, a variety of features have been suggested to be involved in their breakage and mutagenesis. Firstly, it has been proposed that specific DNA features, such as AT-richness or increased flexibility of the double helix could cause difficult-to-replicate secondary structures (14–17). Secondly, CFSs are generally late replicating and this was suggested to make them specifically vulnerable to replication stress (18). In line with this idea, it was also shown that some CFSs have a scarcity of replica-

[*]To whom correspondence should be addressed. Tel: +45 3533 0444; Fax: +45 3532 2128; Email: vibe@bio.ku.dk
Present address: Constanze Pentzold, Institute of Human Genetics, Jena University Hospital, DE-07747 Jena, Germany.

tion origins (19,20) and consequently they cannot respond to replication stress by firing backup origins, which would make these genomic regions more prone to enter mitosis in an underreplicated state (11). Consistently, CFS fragility varies from one cell type to another, implying that epigenetic features rather than the DNA sequence *per se* is causing fragility (9). Finally, many CFSs coincide with large transcribed genes, where potential collisions between replication and transcription machineries can delay replication (21–23). ERFSs, on the other hand, are early replicating, gene dense regions, where high transcription activity of several genes may sensitize the region to replication stress (8).

FANCD2 has an important role in CFS maintenance (24,25) and localizes to CFSs during mitosis in human cells even in the absence of breakage (12,13). FANCD2 prevents replication fork stalling at the AT-rich cores of FRA16D (housing the *WWOX* gene) and FRA6E (housing the *PARK2* gene), and specific examination of FRA16D also shows that FANCD2 deficiency leads to accumulation of RNA-DNA hybrids at this CFS (25). Moreover, FANCD2 is a key component of the Fanconi Anemia pathway that protects cells against genotoxic agents such as DNA interstrand cross-linkers and aldehydes (26).

Traditionally, CFSs have been defined and mapped cytogenetically (11,27) and are mainly characterized in humans, other primates and mice (9,22,28–30). Here, we performed ChIP-seq of FANCD2 from cells subjected to replication stress to map fragile sites genome-wide in the avian cell line DT40. By cytogenetic analyses using fluorescence *in situ* hybridization (FISH) we confirm that the most significant peaks from our ChIP-seq are *bona fide* CFSs. Moreover, we identify a number of gene-dense highly expressed regions likely corresponding to DT40 ERFSs. Additionally, we have developed a fluctuation assay to estimate the site-specific mutation rate at a CFS.

This study provides a genome-wide map of CFSs and putative ERFSs, plotted together with replication timing and transcription level as well as other features suggested to contribute to CFS fragility. We do not find significant correlation with any of the tested sequence features formerly suggested to induce fragility, but our results support that transcription is involved in fragility at both CFSs and ERFSs. Specifically, transcribed genes over 500 kb always coincide with CFSs. Moreover, we report the conservation of long intron sizes in CFSs within the two major branches of the amniotes – the synapsids to become mammals and the sauropsids to become birds and reptiles. Given the fact that these two branches diverged more than 300 million years ago, the conservation suggests that CFSs and the large size of the genes that host CFSs have biological significance yet to be uncovered.

MATERIALS AND METHODS

Cell culture and transfection

DT40 cells were maintained in RPMI 1640 glutamax medium supplemented with 50 μM ß-mercaptoethanol, 10% fetal calf serum, and 1% chicken serum at 37°C with 5% CO_2. Linearized targeting constructs were transfected into DT40 cells by electroporation (Gene Pulser, BioRad). Cell lines used in this study are listed in supplementary methods.

Generation of targeting constructs for *FANCD2*, *PARK2* and *OVAL*

The construct for GS-tagging (protein G-streptavidin binding peptide/tandem affinity purification tag) of endogenous chicken FANCD2 (*FANCD2*, Gene ID: 415935) was generated by first, amplifying the 3′ homology arm using primers VO300 and VO301; and the 5′ homology arm using the primers VO302 and VO303. Next, the 3’ homology arm was cloned into pBluescript SK+ (Agilent Technologies) as a SpeI/NotI fragment. Then, a sequence encoding an N-terminal GS-tag (31) was inserted as a BamHI/XbaI fragment. Thereafter, the 5′ arm was inserted as a SalI/BamHI fragment and finally, the resistance cassette was inserted as a BamHI fragment. The GS targeting construct was linearized with NotI before transfection.

The *BSR*, *PURO*, and *NEO* resistance genes were excised from pLOX-BSR, pLOX-PURO and pLOX-NEO, respectively (32) using BamHI.

For assembly of the *PARK2* (*PARK2*, Gene ID: 421577) knock-in construct (pCP2), the 5′ and 3′ homology arms were PCR amplified using primer pairs CP43-CP44 and CP45-CP46, respectively. Then they were individually subcloned into pCR® 2.1-TOPO® TA vector (Invitrogen) in the listed order using restriction sites KpnI–NotI or NotI–XhoI, respectively. The HyTK cassette was excised by NotI from the original vector (kindly provided by Christine Farr) (33). The *PARK2* targeting construct was linearized with Eam1105I (AhdI) before transfection.

The vector for targeting the chicken ovalbumin gene locus 2 (*OVAL*, Gene ID: 396058) was assembled by first subcloning the 3′ homology arm in the XhoI/KpnI site of pBlueScript SK+ (Fermentas GmbH). The 5′ homology arm was subcloned in the NotI/SpeI site of pLOX-PURO (32). Thereafter, the 3′ and 5′ homology arms were assembled into one vector by subcloning a KpnI/ClaI fragment to generate pML27. The *HyTK* counterselection gene was XhoI-adapted by PCR, using the original vector (33) as template. To finally insert the *HyTK* counterselection gene between the *OVAL* homology arms, the adapted *HyTK* gene was cloned into the XhoI site of pML27, resulting in the final targeting vector pCP9. The *OVAL* targeting construct was linearized with Eam1105I (AhdI) before transfection.

Plasmids and primers used in this study are listed in supplemental methods, respectively.

Chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq)

DT40 WT and FANCD2 $^{GS/GS}$ cell lines were treated with APH (0.5 μM; Sigma) for 16 h. For exp:G2/M, cells were arrested in mitosis with colcemid (0.1 μg/ml; KaryoMax, Gibco). For mitotic enrichment by centrifugal elutriation (34), cells were drawn into a chamber in a Beckmann Coulter™Avanti™ J-20 centrifuge. 150 ml fractions were collected at a flow rate of 38 ml/min at 4000 rpm and rotor speed reducing intervals of 250 rpm. Collected fractions and control cells were spun down and resuspended in PBS. Fractions were immediately fixed for 10 min with 1% formaldehyde (Sigma) for DNA-protein crosslinking. The reaction was stopped by adding glycine (SERVA) to a final concentration of 0.3 M. Each sample was analysed for cell density and viability, used for immunostaining and microscopy, as well as for FACS analysis (shown in Supplementary Figure S1) to identify mitotic fractions in elutriated samples. Cells were washed with TBS (10 mM TrisCl, 150 mM NaCl, pH 7.5) and nuclei isolated by repeated washes in MC lysis-buffer (10 mM TrisCl, 150 mM NaCl, 3 mM $MgCl_2$, 0.5% Igepal CA-630 (Sigma), pH 7.5). The chromatin pellet was snap frozen in liquid nitrogen and thereafter dissolved in FA lysis buffer (50 mM HEPES (Sigma), 150 mM NaCl, 1 mM EDTA (Sigma), 1% Triton X-100 (Sigma), 0.1% sodium deoxycholate (Merck), 0.1% SDS (J.T.Baker), pH 7.5) with a protease inhibitors tablet (Roche). DNA fragmentation was obtained by sonication with the Bioruptor (Diagenode) in 50 cycles of each 10 s pulse (∼85 W) and 30 sec pause. DNA fragmentation was confirmed by gel electrophoresis after reversion of the DNA-protein crosslink by incubation at 95°C for 15 min. FANCD2-bound DNA was pulled down by Pierce Streptavidin UltraLink® Resin acrylamide-based beads (ThermoScientific) equilibrated in FA lysis buffer. The supernatant was removed and the bead mix repeatedly washed in FA lysis buffer with protease inhibitors, followed by each one wash step in FA lysis buffer containing 0.5 M NaCl, in ChIP wash buffer (10 mM TrisCl, 0.25 mM LiCl, 1 mM EDTA (Sigma), 0.5% Igepal CA-630 (Sigma), 0.5% sodium deoxycholate (Merck), pH 8.0), TE buffer and eluted with ChIP elution buffer (50 mM TrisCl, 10 mM EDTA (Sigma), 1% SDS (J.T.Baker), pH 7.5). 1.5 μg/μl Proteinase K (Roche) was added to the supernatant. After DNA-protein crosslink reversal DNA was purified using the QIAquick PCR purification kit (Qiagen) following manufacturer´s protocol. Samples were eluted on gel extraction filters (Fermentas) with TE buffer. Libraries for ChIP-seq were generated using NEBnext® Quick DNA Library Prep Master Mix Set for 454™ (neb #E6090), using MinElute (Qiagen) purification beads, and Illumina inPE adaptors. Library indexing and PCR enrichment (20 cycles) was accomplished according to manufacturer´s instructions using Phusion® high-fidelity DNA polymerase, Illumina primers inPE1.0, inPE2.0 and Illumina index primers. Size selection was validated on a Bioanalyzer (Agilent Technologies) and subsequently sequenced using GAIIx from the Illumina sequencing platform. Exp:async was performed following the same protocol but leaving out colcemid treatment and mitotic enrichment by elutriation and sequenced using the HiSeq2000 (Illumina sequencing platform).

Metaphase spreads

Metaphase spreads were prepared essentially as described (19,35). Briefly, cells in exponential growth were incubated either with or without 0.5 μM APH (Sigma) over night. Cells were exposed to 0.1 μg/ml colcemid (KaryoMax, Gibco) for 3 h. Pelleted cells were gently resuspended in pre-warmed (37°C) hypotonic buffer (FBS, 75 mM KCl, H_2O; 1:1:5; Gibco). Fixation buffer (acetic acid, Ethanol; 1:3; Sigma, CCS Healthcare AB) was added to the cells. After centrifugation, cells were resuspended in fresh fixation buffer and kept at –20°C for at least 16 h before spreading on glass slides.

Fluorescence *in situ* hybridization: FISH

FISH was carried out following the methods described by (35,36) with minor modifications. BACs from the CHORI-261 library listed in supplementary methods were used for preparation of labeled probes. Bacterial strains containing the BACs were spread on LB agar plates containing 12.5 μg/ml chloramphenicol (Sigma). BACs were extracted using the NucleoBond® Xtra Midi Plus kit (Macherey-Nagel) according to manufacturer's instructions.

Probes were labeled with biotin or digoxigenin, using bioprime DNA labeling system (Invitrogen™) or DIG DNA labeling mix x10 (Roche), respectively. For both probes, labeling was carried out according to Invitrogen´s manual using 250 ng DNA as template. Probes were purified on IllustraProbequant G-50 Micro columns (GE Healthcare) following manufacturer´s protocol.

Metaphase spreads were prepared as described above. Slides were hybridized with labeled probes. For immunolabeling, first, streptavidin-Alexa Fluor® 555 (1:200; LifeTechnologies), second biotinylated rabbit α-streptavidin (1:266; Rockland) or mouse α-digoxigenin FITC (1:50; abcam), and third streptavidin-Alexa Fluor® 555 (1:200) or goat α-mouse Alexa Fluor® 488 (1:50; Invitrogen™) were used. Finally, slides were washed in PBS and coverslips were mounted with Vectashield® mounting medium (Vector Laboratories) containing DAPI (1.5 μg/ml; Sigma) for counterstaining of DNA.

Images were acquired with oil immersion on a widefield microscope (AxioImager Z1; Carl Zeiss) equipped with a 100× objective lens (Plan Apochromat, NA 1.4; Carl Zeiss), a cooled CCD camera (Orca-ER; Hamamatsu Photonics), differential interference contrast (DIC), and an illumination source (HXP120C; Carl Zeiss). Three Z-planes with 0.3 μm spacing were imaged. Metaphase spreads were analyzed for breaks or gaps colocalizing with the FISH probes.

RNA-seq

Total RNA was extracted from asynchronously growing DT40 cultures using trizol (Sigma-Aldrich) according to manufacturer's protocol. Strand-specific mRNA-seq libraries for the Illumina platform were generated and sequenced at BaseClear BV (BaseClear, Leiden, The Netherlands). Reads from RNA-seq experiments were mapped with bwa's 'mem' algorithm (37) using default settings to the galgal5 version of the chicken genome downloaded from ensembl release 86. Gene RNA-seq depths were computed using samtools 'bedcov' (38).

ChIP-seq peak calling

Reads from ChIP experiments were mapped with bwa's 'mem' algorithm (39) using default settings to the galgal5 version of the chicken genome downloaded from ensembl release 86. ChIP peaks were called using DROMPA2 (40) using a window size of 100 000 bp and a smoothing window of 200 000 bp. DROMPA2's internal Poisson test was disregarded by setting its p-value cutoff to 1. The FANCD2-ChIP over Control-ChIP enrichment binomial test p-value cutoff was left to the default 0.01 for experiment 1 and set to 0.025 for experiment 2 to account for the weaker signal in the latter.

Circos plots and features

DNA sequence-based properties for the galgal5 genome (AT-content, TwistFlex) were computed by counting mono-, di- and trinucleotide frequencies over 100 kb windows using a simple perl script. Repeat content was similarly gauged by counting repeat-masked Ns over 100 kb windows in the downloaded version of the genome. Counts were converted to Z-scores before plotting for normalisation. Chromosome replication timings were taken from an earlier study (41) and adapted to galgal5 using the hgLiftover tool (42).

Windows corresponding to called peaks and experimentally determined fragile sites were checked for enrichment of any properties by employing linear discriminant analysis (LCA) and support vector machines (SVM).

Circos (39) was used to create chromosome-level plots of all the generated data.

HyTK fluctuation assay to measure mutation rates

PARK2-HyTK or OVAL-HyTK cell lines were grown in 2.5 mg/ml Hygromycin (Gibco) for at least three days. Cells were subjected to limiting dilution (43) in either the absence or presence of APH (0.2 μM; Sigma). After 6–8 days, nine single colonies were picked, both of untreated and APH-treated cells and scaled up. One to three days later, depending on cell density, cells were harvested and dispensed in medium containing 1 μM ganciclovir (Sigma) in 96-well plates (200 μl/well). The density at which cells were plated was tested beforehand in dilution cloning experiments to define the specific dilution factor for each cell line and condition (±APH). For the estimation of the number of mutational events at the *HyTK* locus, the number of revertant colonies (*r*) in 96-well plates was counted about two weeks after plating in ganciclovir. Mutation rates were calculated using the web-based tool ‘FALCOR’ (44) that works with assumptions from Luria-Delbrück fluctuation analysis (45) differentiated in the Ma-Sandri-Sarkar Maximum Likelihood Estimation Method (MSS-MLE). For calculations, the number of revertants (ganciclovir resistant clones, *r*) and the total number of plated cells (N_t) was used. Average mutation rates were determined after three to five independent experiments for each cell line and condition.

Vertebrate comparative genomics

All 204 vertebrate genomes available on NCBI’s FTP site as of November 2016 were downloaded (Supplementary Table S6). The veNOG dataset from the eggNOG orthology database (46) was used to find the corresponding genes across all 204 genomes. Although the original veNOG orthology was based on an older set of vertebrate genomes than the set we downloaded, eggNOG-mapper (47) was used to map the original veNOG groups onto the new genomes. eggNOG-mapper was run using the ‘–target_orthologs one2one’ setting to get as conservative a mapping as possible.

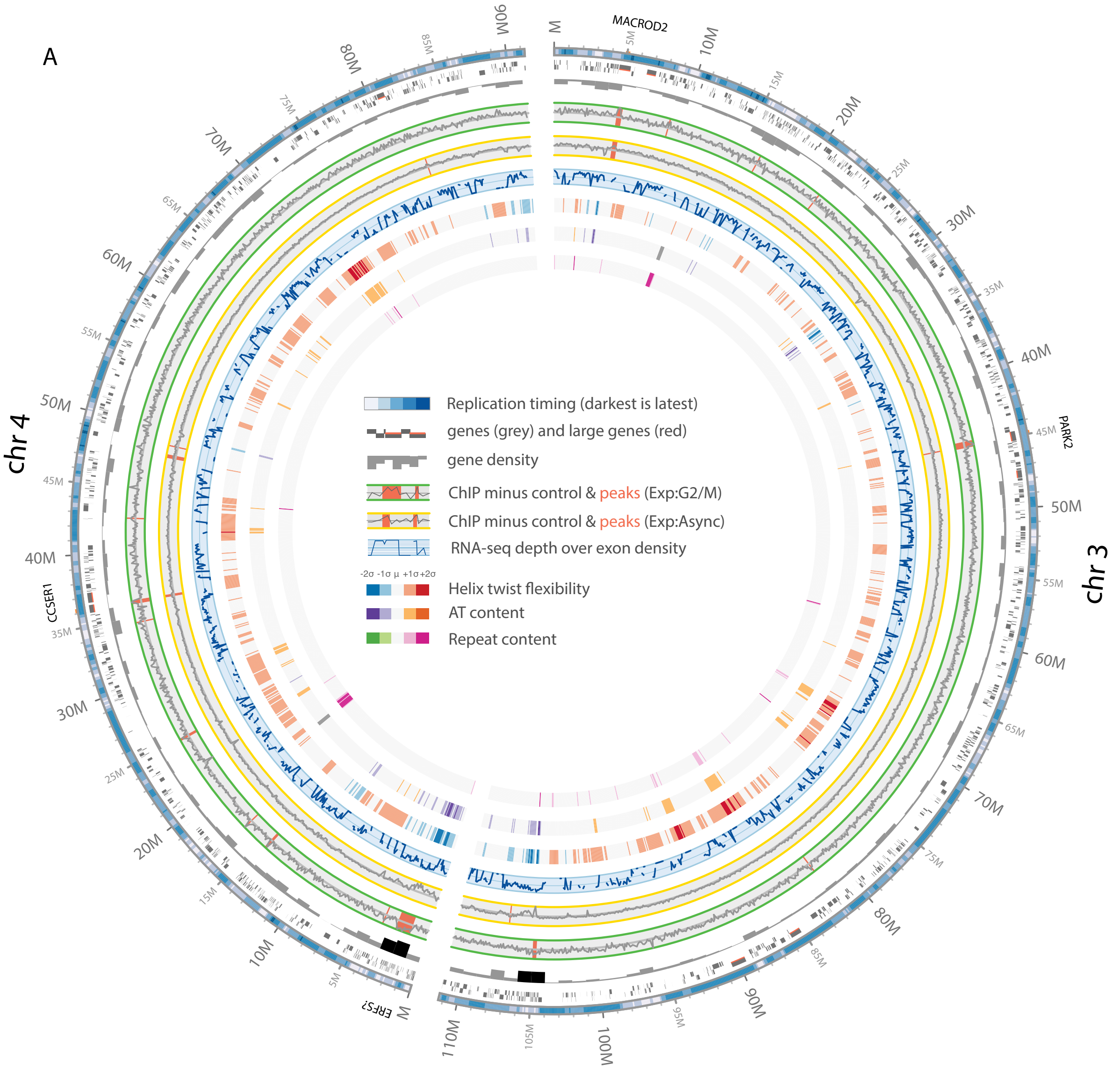
RESULTS

ChIP-seq of FANCD2 identifies putative fragile sites in the avian cell line DT40

The chicken genome is organized into five pairs of autosomal macrochromosomes and one sex macrochromosome (>40 Mb), four pairs of intermediate sized autosomal chromosomes (20–40 Mb), and 28 pairs of autosomal and one sex microchromosome (<10 Mb) (48). Due to the small size of most chromosomes, complete karyotyping and genome-wide cytogenetic mapping of CFSs is not achievable by current techniques.

Human FANCD2 binds to CFSs on mitotic chromosomes in response to replication stress (12,13) and similar localization of Venus-tagged FANCD2 to mitotic chromosomes has been observed in chicken DT40 cells (Supplementary Figure S2A and (49)). To confirm this, we tagged endogenous *FANCD2* on both alleles in DT40 cells with a tandem-protein G-streptavidin tag (GS-tag) (31) to obtain the cell line $FANCD2^{GS/GS}$(Supplementary Figure S2B). The GS-tagged FANCD2 forms spontaneous foci on mitotic chromosomes, which are further induced by low doses of the replication inhibitor aphidicolin (APH) (Supplementary Figure S2C). Moreover, the biological functionality of the GS- and Venus-tagged FANCD2 was confirmed by lack of sensitivity to the interstrand cross-linker cisplatin in the $FANCD2^{GS/GS}$ and $FANCD2^{Venus/Venus}$ cell lines (Supplementary Figure S2D).

To map fragile sites in DT40, we conducted ChIP-seq of GS-tagged FANCD2. $FANCD2^{GS/GS}$ as well as unmodified DT40 cells, were subjected to mild replication inhibition (0.5 μM APH). In one experiment (exp:G2/M), G2/M cells were enriched by centrifugal elutriation before ChIP-seq. In a second experiment (exp:async), asynchronous APH-treated cells were subjected to ChIP-seq. Reads from the two experiments were mapped to the chicken genome. From these experiments, we identified 96 and 82 significant FANCD2 binding sites for exp:G2/M and exp:async, respectively (Supplementary Tables S1 and S2). The results from the two experiments are presented in circos plots of the chicken genome (Figure 1A (chromosomes 3 and 4) and Supplementary Figure S3 (whole genome)). In 12 out of the 96 and 18 out of 82 (for exp:G2/M and exp:async, respectively), the FANCD2 ChIP-seq peaks coincide with genes spanning more than 200 kb (Supplementary Tables S1 and S2). Strikingly, *PARK2*, *MACROD2, GRID1* and *CCSER1,* all of which span more than 500 kb, overlap with four of the seven most significant peaks in exp:G2/M (Figure 1B). Notably, *MACROD2, CCSER1, PARK2* and *GRID1* were also identified as significant FANCD2 binding sites in exp:async along with the *WWOX* gene, the human homolog of which is located at the human fragile site FRA16D (Supplementary Table S2) (50). Circos plots of specific genomic regions hosting *PARK2, MACROD2, GRID1* are shown in Supplementary Figure S4. Peaks in exp:G2/M were most significant (Supplementary Table S1), therefore the following analyses focus mainly on peaks from this experiment.



B

chromosome	Start position	End position	logp (IP/control)	Overlapping large gene (more than 200kb)	gene size (bases)
chr3	44600000	45099999	15.95	PARK2	677633
chr3	4900000	5299999	15.05	MACROD2	795578
chrZ	0	799999	13.07		
chr6	1700000	2199999	11.8	GRID1	517606
chrW	3800000	4099999	10.51		
chr18	9300000	10999999	10.2		
chr4	36000000	36299999	10.11	CCSER1	548799
chrZ	8700000	9299999	8.9		
chrZ	2700000	2899999	8.17	SETBP1	247219
chr27	5200000	5599999	7.8		

Figure 1. Genome-wide FANCD2 binding analysis. (**A**) Circos plot of *Gallus gallus* chromosomes 3 and 4 indicating replication timing, gene position, gene density, FANCD2 ChIP-seq results, transcription levels and DNA sequence features. Replication timing obtained by (41) is indicated by blue shading where darker is later (level 9, outermost circle). Position of genes in forward or reverse orientation is shown as gray bars, above or below center, respectively (level 8). Genes larger than 500 kb are marked with a red line at the center (level 8). Gene density per megabase window is indicated with gray bars. The average gene density in the chicken genome is 16.6 genes per megabase with a standard deviation (σ) of 13.1. Regions with a gene density of more than

The most significant FANCD2 binding sites form gaps and breaks on metaphase chromosomes in response to replication stress

To investigate whether FANCD2 binding sites correspond to *bona fide* CFSs, we employed FISH on metaphase spreads at three of the seven most significant peak regions from FANCD2 ChIP-seq as well as two control sites. The sites were chosen because they were located on macrochromosomes and therefore suitable for FISH analysis. A site was considered to have a break/gap if the chromosomal region was not stained by DAPI at the FISH signal. For the three sampled FANCD2 binding sites (corresponding to the genomic loci for the genes *PARK2*, *MACROD2* and *CCSER1*), a significant increase in breaks/gaps was observed in response to APH treatment (Figure 2A and B). Control regions (*RNF8* and *BCL2*) did not exhibit increased breakage when treated with APH (Figure 2A). Taken together, these results reveal that FANCD2 ChIP-seq from APH-treated cells identifies *bona fide* highly fragile CFSs in avian DT40 cells.

FANCD2 ChIP-seq from G2/M cells also identifies putative ERFSs

Curiously, many of the peaks identified in exp:G2/M did not coincide with large genes. We thus sought for other features that could explain FANCD2 localization to these genomic loci. Sequence characteristics, such as AT content, twist-flexibility or simple repeats, have been suggested to contribute to the fragility at CFSs and we therefore included these in the circos plots (Figure 1 and Supplementary Figure S3). The correlation of these and other DNA features with the identified FANCD2 binding sites was calculated by discriminative analysis in order to search for primary sequence similarities among all FANCD2 binding regions. High CpG content displays weak correlation with the FANCD2 binding sites compared to non-enriched sites (Supplementary Table S3). However, none of the sequence-derived features were found to correlate significantly with the peaks and even when used in combination, the weak correlations did not yield enough discriminatory power for computational prediction of peaks (Supplementary Table S3).

We did however find that many peaks identified in exp:G2/M were localized in gene dense regions that might correspond to ERFSs (Figure 1A and B). ERFSs are early replicating whereas CFSs are generally thought to be late replicating (8,11,23). To correlate replication timing in DT40 with CFSs and putative ERFSs, existing data on replication timing in DT40 (41) was also plotted in the circos plots. The putative ERFS on chromosome 18 ranked number 6 in epx:G2/M was also included in the detailed circus plot (Supplementary Figure S4). This revealed that the large genes *MACROD2*, *PARK2*, *GRID1* and *CCSER1* were mid-to-late replicating, and the putative ERFSs were indeed early replicating (Figure 1A and Supplementary Figure S3). To reveal whether the putative ERFSs were also highly transcribed, which is a defining feature of *bona fide* ERFSs (8), we conducted RNA-seq to map and quantify transcription in the two DT40 cell lines, the parental DT40 and the derived *FANCD2*$^{GS/GS}$ cell lines. The two cell lines did not differ in their pattern of transcription. Thus, the combined data is shown as one track in the circos plots (Figure 1A and Supplementary Figures S3 and S4).

A heatmap of all significant peaks from the exp:G2/M was generated using ClustVis (51) based on RNA-seq, replication timing, and gene density. ChIP-seq values from exp:async was also included to reveal differences between exp:G2/M and exp:async results (Figure 3). Exp:G2/M peaks largely fall into two clusters with different genomic characteristics. Cluster A scores high for FANCD2 ChIP in exp:async (35 out of 36), has low gene expression levels (24/36), is late replicating (36/36), and has low gene density (27/36). The other cluster, cluster B, scores low for FANCD2 ChIP in exp:async (62/62), has high expression levels (62/62), is early replicating (54/62), and is gene dense (58/62). The early replicating cluster B includes the histone locus, which has previously been identified as an ERFS (8,52).

Transcription of genes above 500 kb strongly correlates with their fragility

To understand why only certain very long genes are fragile, we further analyzed the largest annotated genes. First, we plotted the size of genes spanning >200 kb against their replication timing (Figure 4A). This plot shows that in DT40 cells, genes >500 kb are generally late replicating (values < 0 in Figure 4A) regardless of their expression status, with *SASH1* and *PARK2* being the only exceptions. 90% of the late replicating genes >500 kb are not fragile, showing that late replication timing *per se* does not explain fragility at the large genes (Table 1).

To reveal how transcription of large genes relates to their fragility, we carefully inspected the 33 largest genes, which span from 514 to 1032 kb. *CCSER1, MACROD2* and *PARK2* (marked with yellow in Table 1), which were identified as potential CFSs, are indeed transcribed. The RNA levels of these transcripts are relatively low but the RNA-seq reads span >500 kb with even distributions across the gene bodies (Table 1 and Figure 4B). The same pattern is seen for *GRID1* (marked with yellow in Table 1) though its transcription level is even lower and reads span only ∼450 kb (Table 1 and Figure 4B). Given the low level of the *GRID1* transcript, it is difficult to rule out that the actual

16.6 +2*σ are black (level 7). The ChIP-seq data is represented in 100 kb windows showing the –log(p) values from two FANCD2 ChIP-seq experiments (exp:G2/M, marked by green lines) and (exp:async, marked by yellow lines) where cells treated with APH have or have not been subjected to elutriation before ChIP of FANCD2, at levels 6 and 5, respectively. Significant peaks identified using the DROMPA2 peak finder are shaded in red (40). RNA-seq results are plotted as RNA-seq depth over exon density at level 4. Twist angle flexibility (level 3), AT content (level 2) and repeat content (level 1, innermost circle) are indicated by color codes showing deviations from mean as indicated. Positions of *MACROD2*, *PARK2*, *GRID1*, *CCSER1* and a putative ERFS are indicated. (**B**) The 10 most significant peaks from exp:G2/M listed with the position in the genome, the -log(p) value and name and size of overlapping genes >200 kb.

Table 1. Transcription status and replication timing in DT40 of gallus gallus genes larger than 500 kb

geneID	Gene size (bases)	Gene name	Size of transcribed region according to RNAseq data (bases)	Replication timing (au)	RNAseq (total number of reads)*	Gene within peak in exp:G2-M	Gene within peak in exp:async
XP_015140406	1032419	CSMD1		-1.4	0	no	no
NP_990630	993455	DMD	51	-1.6	1	no	no
XP_015154014	914239	ROBO2	51	-1.07	1	no	no
XP_015142894	892004	NRXN3	674026	-1.56	2	no	no
XP_015131476	838904	CNTNAP2	776543	-1.46	21	no	no
XP_015138910	795578	MACROD2	795502	-1.47	132	yes	yes
XP_015140331	773622	LOC100858320	32	-1.42	2	no	no
XP_414942	744781	RBFOX1	254165	-1.6	738	no	no
XP_015130158	736621	GPC6	465449	-1.49	84	no	no
XP_015131225	722801	MAGI2	689325	-1.19	6	no	no
XP_416995	700805	PCDH9	51	-1.33	1	no	no
XP_004937134	695511	CELF4		-1.55	0	no	no
XP_015132086	694737	GRID2	280	-1.66	286	no	no
XP_015137268	691855	RBMS3		-1.83	0	no	no
XP_419615	677633	PARK2	677096	0.04	81	yes	yes
XP_015151429	660055	AUTS2	12027	-1.63	3	no	yes
XP_015153869	648260	CADM2	192326	-1.49	11	no	no
XP_015133503	643404	NRXN1	39	-1.59	1	no	no
XP_015137094	639335	PTPRN2	600385	-1.63	3	no	no
XP_015135588	626368	DLG2		-1.07	0	no	no
XP_004938402	617039	ROBO1		-1.49	0	no	no
XP_425563	616130	IL1RAPL1		-1.57	0	no	no
XP_004938641	613931	GPC5	419240	-1.49	41	no	no
XP_015143131	608598	NPAS3	500219	-1.45	3	no	no
XP_015145376	577956	LRP1B		-1.63	0	no	no
XP_418391	566340	CSMD3	1429	-1.39	581	no	no
XP_015154652	549205	LOC769232	84441	-1.84	7	no	no
XP_004941092	548799	CCSER1	548785	-1.71	171	yes	yes
XP_015146714	527557	AGBL4	334590	-1.85	14	no	no
XP_004935653	526499	SASH1	504940	0.99	27	no	no
XP_015141352	525168	SORCS2		-1.22	0	no	no
XP_015143515	517606	GRID1	424527	-1.53	42	yes	yes
XP_015137743	514384	CTNND2	252639	-0.23	420	no	no

*Total number of mapped reads is 17725385. All annotated genes larger than 500 kb are listed. *Yellow*, fragile genes. *Grey*, non-fragile genes with no expression. *Purple*, non-fragile genes with uneven read distribution. *Blue*, non-fragile genes with even low read distribution. *Pink*, non-fragile gene with high expression in a 252639 large region. See text for further details.

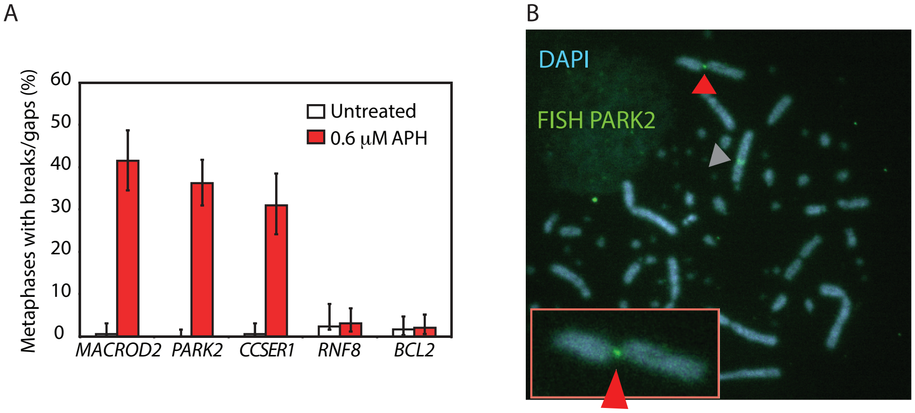


Figure 2. Peak regions from FANCD2 ChIP-seq form breaks on metaphase chromosomes after APH treatment. (**A**) Quantification of breaks or gaps at the indicated loci with or without 0.6 μM APH treatment for 16 h. *BCL2* and *RNF8* are control loci, which were not bound by FANCD2 in the ChIP-seq experiment. At least 170 metaphases were counted for each locus. (**B**) Representative image of metaphase spread from DT40 cells treated with 0.6 μM APH for 16 h. The FISH probe hybridizes to *PARK2*, which is broken (close up image and red arrowhead). One allele of *PARK2* is not broken (grey arrowhead).

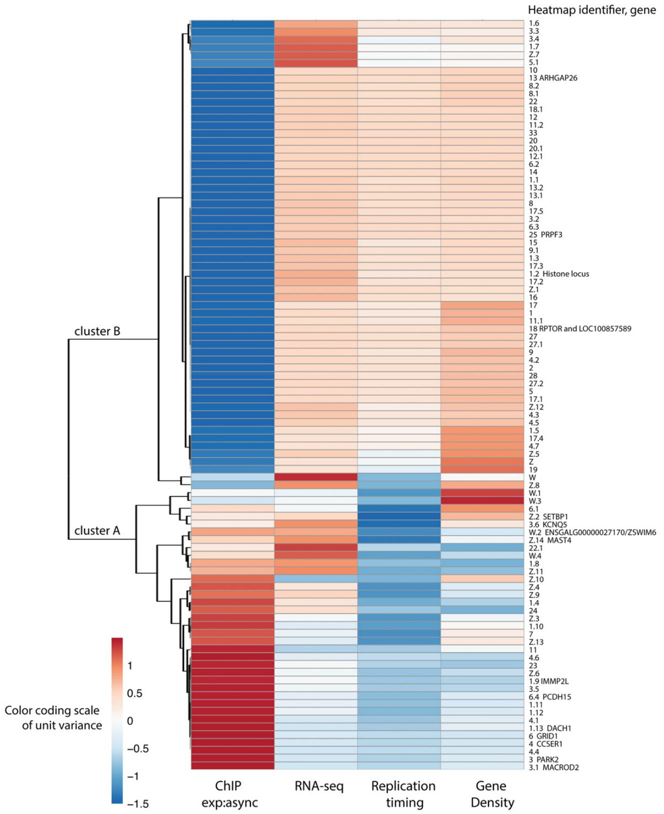


Figure 3. FANCD2 ChIP-seq peaks from exp:G2/M form two distinct clusters based on exp:async, transcription, replication timing and gene density. Heatmap showing multivariate clustering of significant peaks from exp:G2/M using peak values from ChIP-seq exp:async, RNA-seq, replication timing and gene density. Cluster A includes putative CFSs, which are characterized by FANCD2 binding, low transcription, late replication and low gene density. Cluster B includes putative ERFSs, which are characterized by FANCD2 binding only in G2/M, high transcription, early replication and high gene density. Rows are centered; unit variance scaling is applied to rows. Red colors indicate a high relative value and blue colors indicate a low relative value (as indicated by the red/blue heatmap unit variance color scale). Rows are clustered using correlation distance and average linkage. 96 rows, 4 columns.

transcribed region could be longer. The mRNA levels observed for the large fragile genes are low compared to other expressed genes. For comparison; of the total 17 725 385 mapped reads, 72 614 reads mapped to the highly expressed *ACTB* gene, whereas 149 reads mapped to the poorly expressed *POLK* gene. Between 42 and 171 reads mapped to the large fragile genes *CCSER1, MACROD2*, *PARK2* and *GRID1*.

17 out of the 33 largest genes have no or very low expression (corresponding to zero to three RNA-seq reads in the entire gene, marked with gray in Table 1). None of these are found in peak regions in FANCD2 ChIP-seq exp:G2/M, suggesting that transcription of the large genes is required to induce fragility at the gene. As an exception to this pattern, one long gene, *AUTS2*, with a very low expression level (three reads) is located in a peak region in exp:async.

Certain very long genes apparently have a high transcription rate according to the RNA-seq data, but are not identified by FANCD2 ChIP-seq. These include *RBFOX1*, *GPC6, CSMD3* and *GRID2* and others (marked with purple in Table 1). However, most or all RNA-seq reads map to a specific part of the gene showing that the high transcription level does not reflect transcription of the entire gene (Figure 4C).

Intriguingly, some of the large genes, including *GPC5*, *CNTNAP2*, *AGBL4* and *SASH1* (marked with blue in Table 1), seem to escape fragility as detected by FANCD2 ChIP-seq even though considerable levels of very long transcripts are produced from these genes (Figure 4D). It is worth noting that *SASH1* actually is the only gene over 500 kb that is clearly early replicated (Figure 4A). All of these genes have lower transcript levels than *CCSER1*, *MACROD2*, *PARK2* and *GRID1*, indicating that very low transcription levels of long genes may be tolerated without causing fragility to the genes.

Finally, the gene *CTNND2* displays a slightly different pattern having relatively high mRNA levels in a 250 kb window (Table 1 and Figure 4E). This indicates that robust transcription of transcripts slightly longer than 200 kb does

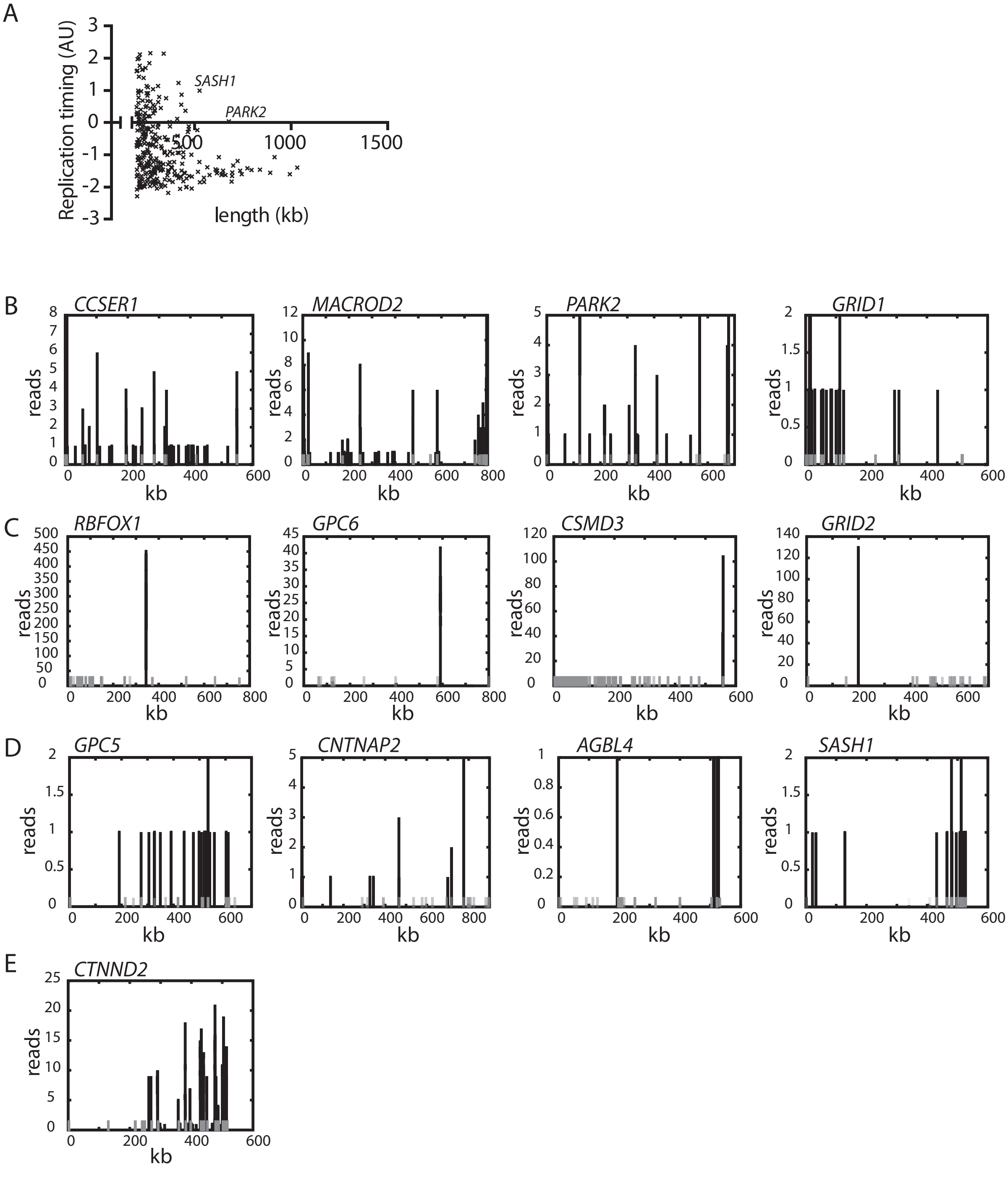


Figure 4. Very long transcripts are generated at CFS genes. (**A**) Long genes replicate late in the cell cycle. Graph shows the replication timing of genes >200 kb. Gene size (kb) given on the X-axis is plotted against replication timing in arbitrary units (AU) on the Y-axis with highest values corresponding to early replication. (**B**) Fragility of long genes correlate with transcription. Plots show the absolute RNA-seq read numbers in large fragile genes. (**C**–**E**) Non-fragile genes exhibit low or partial expression. Plot shows the absolute RNA-seq read numbers from large non-fragile genes. The X-axis represents the gene including introns. Exons are indicated by grey pins on the X-axis. Position of the reads (black) are indicated relative to the start of the coding sequence, X = 0. The Y-axis shows the number of reads. The total number of mapped reads was 17 725 385.

not cause replication problems detected by FANCD2 ChIP-seq.

In summary, FANCD2 ChIP-seq from G2/M cells identifies two types of genomic regions, one type includes CFSs coinciding with extremely large transcribed genes that are mainly late replicating. The other type is reminiscent of ERFSs, consisting of gene dense and early replicating regions. Curiously, the putative ERFSs identified in the exp:G2/M experiment score very low for FANCD2 binding in exp:async, whereas putative CFSs have relatively high scores in both experiments.

Replication stress is highly mutagenic to the CFS *PARK2*

CFSs are hotspots for copy number variations (CNVs) and genomic rearrangements (9,11,53,54), but so far the actual mutation rate at a CFS has not been determined. To accomplish this, we developed a fluctuation assay exploiting the fusion gene Hygromycin phosphotransferase-Thymidine Kinase (*HyTK*), allowing for both positive and negative selection using hygromycin and ganciclovir, respectively (33,55). *HyTK* was integrated in the fifth intron of the highly fragile *PARK2* gene (Figure 5A) as well as at a control site, *OVAL*, which is a transcriptionally inactive locus (56). The resulting cell lines are referred to as *PARK2-HyTK* and *OVAL-HyTK*, respectively.

The mutation rate at the *PARK2* locus was tested using four independently derived clones in the absence or presence of APH. All clones responded to APH with a significant increase in mutation rates in the range of around 50 to more than 300 mutations per million cell divisions (Figure 5B). In contrast, the mutation rate at the *OVAL* locus did not significantly increase in response to APH (Figure 5B).

An interclonal variation in mutation rate was observed in the *PARK2-HyTK* clones (Figure 5B). To test whether this interclonal difference in mutation rate correlates with a difference in fragility of the *PARK2* locus in the different clones, we performed FISH to analyze breaks at *PARK2* in *PARK2-HyTK* cl. 1 and cl. 2 (Supplementary Figure S5). Although there was a trend towards increased fragility correlating with elevated mutation rate in cl. 1, there was no significant difference in fragility at *PARK2* between *PARK2-HyTK* cl. 1 and cl. 2 (Supplementary Figure S5). In conclusion, replication stress is highly mutagenic to the identified CFS harboring *PARK2*.

APH induces deletions at the CFS *PARK2*

To classify the mutational events and to confirm the occurrence of mutations in the ganciclovir resistant clones isolated from the HyTK assay, we employed Southern blot, PCR and Sanger sequencing on genomic DNA from a number of resistant PARK2-HyTK and OVAL-HyTK clones isolated following untreated or APH-treated conditions (Supplementary Table S4 and Figure S6). 26 out of the 27 ganciclovir resistant clones isolated from APH-treated PARK2-HyTK clones harbored deletions (Figure 5C). One of these harbored a homozygous deletion (P10, Supplementary Figure S6B). The one clone that did not have a deletion had a one base pair insertion in the *HyTK* coding regions as shown by Sanger sequencing (Supplementary Table S4). Thus, we could identify a mutational event in all 27 analyzed clones. In untreated PARK2-HyTK, 9 out of 19 clones had deletions of the *HyTK* cassette (Figure 5C and Supplementary Figure S6C). In four of the remaining clones, a mutation was detected in the *HyTK* cassette. In 4 other clones, no mutations were found in the *HyTK* cassette (two were not sequenced) (Supplementary Table S4).

In the isolated ganciclovir resistant OVAL-HyTK clones, deletion events were less frequent (Figure 5C) and moreover we could not detect any mutations by sequencing of

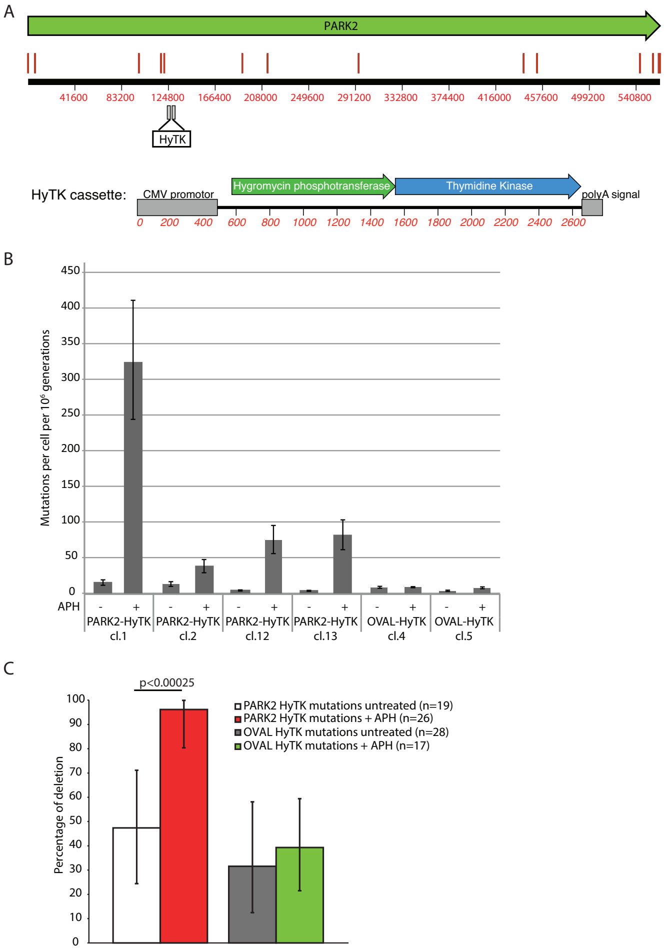


Figure 5. The *PARK2* locus has a high mutation rate in response to APH. (**A**) Top. Map of the *PARK2* genomic locus. The chicken *PARK2* gene is indicated by the thick green arrow. Exons are indicated by red boxes. Ruler indicates position along the gene (bases). Integration site for the *HyTK* cassette is indicated. *HyTK* is not drawn to scale. Bottom. Schematic representation of the *HyTK* cassette, including the CMV promotor, polyA signal and the coding sequence for hygromycin phosphotransferase (*green*) fused to the coding sequencing for HSV1 thymidine kinase (*blue*). Ruler indicates position along the gene (bases). (**B**) Mutation rates at *PARK2* and at the *OVAL* control locus with or without treatment with the replication inhibitor APH. Briefly, single cells were expanded for 10 days in medium with or without 0.2 μM APH. Single colonies were expanded further before dilution-plating in ganciclovir counter-selective medium. The number of colonies appearing is used to calculate mutation rates using the Ma-Sandri-Sarkar Maximum Likelihood Estimator method provided by FALCOR (44). Error bars represent 95% confidence intervals. (**C**) Percentage of deletions in a number of mutants derived from PARK2-HyTK and OVAL-HyTK cell lines, which were untreated or APH-treated. P-value is indicated (exact binomial test).

the coding region of the *HyTK* gene, suggesting that ganciclovir resistance is due to mutation at another genomic location rather than the *OVAL* locus (Supplementary Figure S7 and Table S5).

In conclusion, the gain of ganciclovir resistance by mutation at *PARK2* is mainly driven by deletion events. This faithfully mimics the observation that human CFSs are hotspots for CNVs most of which are large deletions (53).

Large introns of CFS genes have resisted size reduction through vertebrate evolution

The results obtained here demonstrate that expression of extremely large genes gives rise to chromosomal fragility. To add an evolutionary perspective on gene length, we analyzed its distribution within vertebrates. For this purpose we extracted data on orthologous genes from a total of 203 vertebrates, which we divided into the four classes: Bird (Aves), Fish (Chondrichthyes and Osteichthyes), Reptile plus Amphibian (10 species representing Reptilia and 3 species representing Amphibia) and Mammal (Mammalia).

Generally, extremely large genes consist mainly of large introns and code for small proteins, such as the 50 kDa Parkin encoded by *PARK2* spanning more that 670 000 base pairs. Despite the proven susceptibility of extremely large genes to large deletion, evolution has not eliminated the introns of extremely long genes. This is evident from the data as the extremely large genes in *Gallus gallus* were also found to have extremely large orthologs in most other vertebrates, though global differences between species were also detectable. This prompted us to perform a detailed descriptive analysis of intron size for the extremely large genes.

The initial analysis (supplementary report S8) revealed notable differences in intron size distribution between species with birds and fish generally having the smallest introns and mammals having the largest introns. To describe the variation in intron size we fitted the following additive model on a log-scale

$$E(\log_2(L(g,s))) = \alpha_g + \beta_s,$$

where $L(g,s)$ denotes the total intron size of gene g in species s, and where α_g and β_s denote the additive gene effect and species effect, respectively, on the log-scale (base 2). It was found that introns in extremely long genes are generally a factor two to eight smaller in birds and fish than in most mammals. This is compatible with an evolutionary tendency of intron size reduction in birds, the closest living relative of crocodilians (reptiles), which generally have larger introns than birds (57).

The additive model did, however, not fit the data perfectly, with some systematic deviations in its intron size predictions for species within the Bird class (Figure 6, left). The analysis prompted us to split the Bird class into BirdA and BirdB subclasses according to a notable difference in the overall intron size distribution. A better fit (Figure 6, right) was then obtained using a class-gene interaction model,

$$E(\log_2(L(g,s))) = \alpha_{g,c(s)} + \beta_s,$$

where $\alpha_{g,c}$ denotes the combined effect on intron size for gene g in class c, and $c(s)$ denotes the class that species s belongs to.

It is notable from Figure 6 that the large fragile genes *PARK2, CCSER1, MACROD2, GRID1, IMMP2L, DACH1* and *SETBP1* all have larger introns than predicted by the additive model whereas the genes *RPTOR*, *ARHGAP26* and *LOC100857859* have smaller introns than predicted by the additive model. Intriguingly the latter three genes cluster together with putative ERFSs, in contrast to *PARK2, CCSER1, MACROD2, GRID1, IMMP2L, DACH1* and *SETBP1*, which belong to the CFS cluster (Figure 3). Thus, the *Gallus gallus* CFS genes in the BirdB subclass have larger introns than anticipated by the additive model. The class-gene interaction model can correct this lack of model fit, and this shows that *PARK2, CCSER1, MACROD2, GRID1, IMMP2L, DACH1* and *SETBP1* to some extent have resisted the global tendency of intron size reduction in the Bird class despite their fragility, indicating that long introns in these genes might have a beneficial function.

DISCUSSION

Generally, mutations are thought to accumulate gradually with a random distribution throughout the genome. However, certain regions of the genome known as fragile sites are more prone to mutate and frequently undergo large structural rearrangements (2). So far, the underlying cause of CFS breakage seems to be elusive and not restricted to certain features (58). Here we demonstrate that FANCD2 ChIP-seq from G2/M cells subjected to mild replication stress detects CFSs and probably also ERFSs, providing an unbiased genome-wide map of fragile sites in the avian DT40 cell line. FANCD2 accumulation at large genes was observed by ChIP-seq from both asynchronous APH-treated cultures and G2/M-enriched cell cultures, whereas putative ERFSs were only identified in FANCD2 ChIP-seq from the G2/M-enriched cell population.

The identification of putative ERFSs by the exp:G2/M protocol (0.5 μM APH for 16 h followed by 0.1 μg/ml colcemid for 2 h and mitotic enrichment by centrifugal elutriation), but not when elutriation and colcemid treatment is omitted (exp:async) suggests that ERFSs are primarily bound by FANCD2 in G2/M cells in contrast to CFSs, which were identified by both protocols. One possible explanation for this difference could be that FANCD2 is excluded from ERFSs due to their high level of transcription in interphase cells, even if replication is inhibited at the sites. When cells enter mitosis, transcription is shut down (59) allowing FANCD2 to accumulate at the underreplicated ERFSs. Moreover, completion of mitosis might be necessary for resolving problems at ERFSs. By comparison, ERFSs were defined in mouse cells using very high doses of hydroxyurea (8) using ChIP-seq of RPA, BRCA1 and SMC5 from B cells primarily in G1 and early S, indicating that under these conditions ERFSs are recognized by certain repair proteins in S phase. Interestingly, Barlow et al. also show that breaks or gaps on metaphase chromosomes often form at ERFSs strongly suggesting that ERFSs can stay unrepaired from early S to mitosis, but further experimental work will be required to fully understand the conditions that lead to breakage of ERFSs.

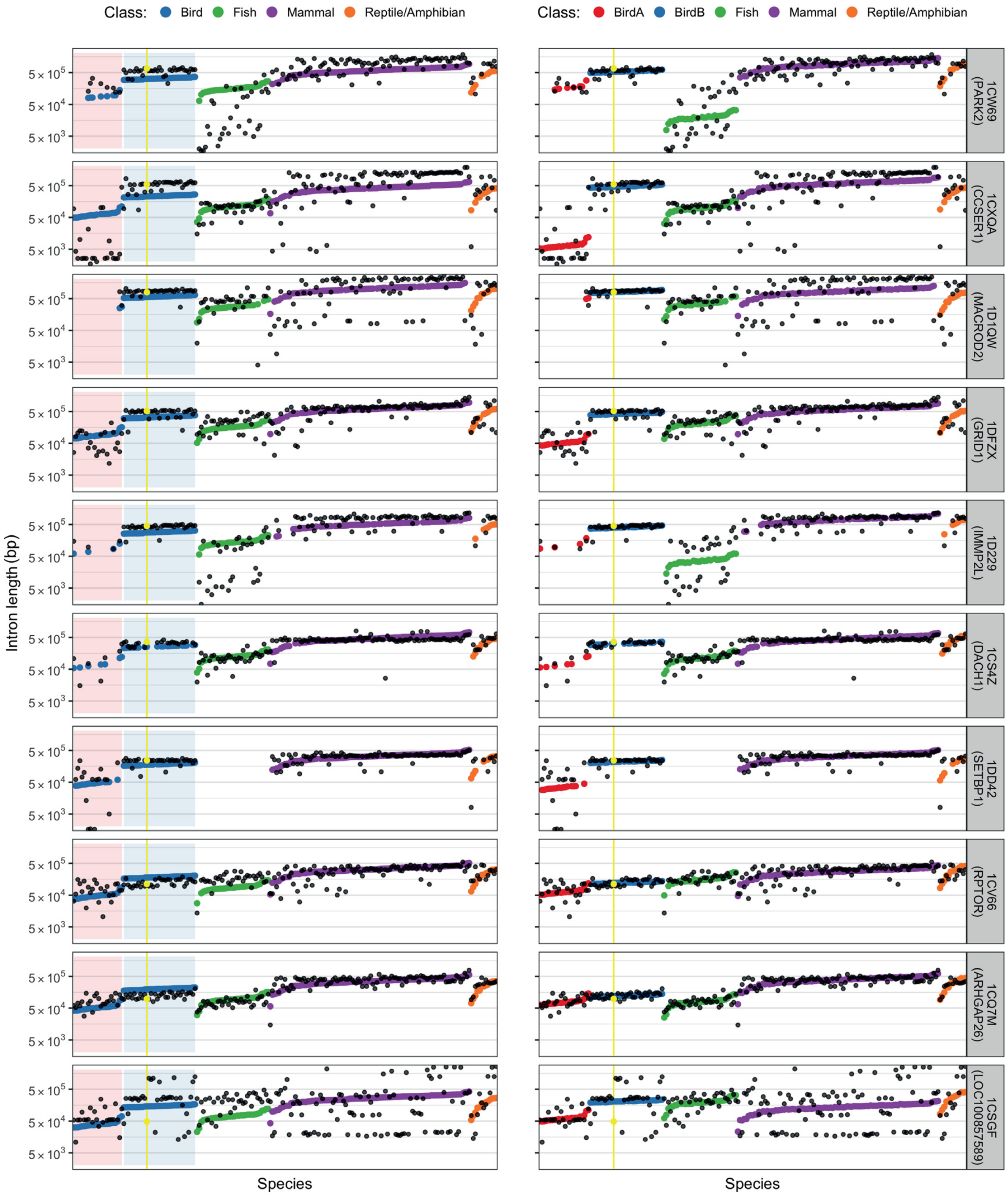


Figure 6. Class specific intron length preservation in long fragile *Gallus gallus* genes. Intron lengths in base pairs (black points) for 10 genes and 203 vertebrates together with fitted values (colored points) for the additive model (left) and the class-gene interaction model (right). Classes and subclasses are indicated by different colors. *Gallus gallus* is marked by yellow. Gene names are given as the ortholog identifier with *Gallus gallus* gene names in parentheses. The subgrouping of Bird identified by the additive model is marked by the shaded background on the left figure.

The lack of complete synteny between gene clusters at ERFSs in mouse and chickens complicates comparison of the putative chicken ERFSs with the previously reported mouse ERFSs (8). However manual inspection of the top ranked putative ERFSs revealed that the putative ERFS on chicken chromosome 18, ranked number 6 in exp:G2/M, is likely corresponding to the mapped ERFS on mouse chromosome 11 position 117 767 080–120 462 465. At least 12 homologs of the 19 mouse genes in this ERFS are present in the putative chicken ERFS reported here. A zoomed in circos plot is also included for this region in Supplementary Figure S4. Given the dependence on transcription, both CFSs and ERFSs are expected to vary between different cell types, which further complicates direct comparison with the ERFSs reported by Barlow et al. and may explain why we not find extensive overlap between chicken DT40 ERFSs and the gene clusters found at activated primary mouse B lymphocyte ERFSs.

Our analysis of CFSs provides compelling evidence that late replication timing per se is not sufficient to induce fragility. Also, sequence features such as AT content do not show any significant correlation with fragile sites. Rather, robust transcription of genes longer than 500 kb in all cases leads to highly significant fragility. The *GRID1* gene also seems to be highly fragile although we could only detect a low abundant transcript of 425 kb from this gene. However, the RNA seq method used here is likely underestimating the actual size of the transcript. Detection of nascent RNA with global run-on sequencing (GRO-seq) (60) or Bru-seq (61) would be useful to determine the minimal sizes of long transcripts that trigger CFS fragility. Moreover based on our data we cannot rule out that very low levels of extremely large transcripts may be tolerated or may cause low levels of fragility that are not detected by FANCD2 ChIP-seq. The notion that large transcription units is the primary cause of CFSs is consistent with previous findings (53). Transcription of extremely long genes poses a particular challenge to dividing cells because the time it takes for RNA polymerase II to transcribe the entire gene may exceed the duration of one cell cycle (21). As a consequence transcription/replication conflicts are inevitable. Such conflicts are efficient triggers of genomic instability (62,63) probably via replication inhibition by so-called R-loops, where nascent RNA leaving the exit pore of RNA pol II hybridizes back to the complementary DNA template to form an RNA–DNA hybrid (64). Consistent with the notion that transcription/replication conflicts are triggers of CFS instability, the direct involvement of R-loops in CFS breakage has been reported (21,25). At the same time transcription of long genes may also indirectly contribute to CFS instability by clearing the transcribed region of replication origins as suggested by Wilson et al. (53). This happens because RNA pol II displaces the pre-replication complexes as it moves along the gene (65–67), which may result in a local scarcity of origins contributing to fragility at CFSs (19).

The mutation rate at CFSs has not previously been determined. To undertake this task, we developed a fluctuation assay, which allowed us to estimate the mutation rate per cell division. The results indicate that as many as 8–10 cells per 100 000 divisions will acquire mutations at the analyzed site in *PARK2*, when replication stress is induced. For comparison mutation rates reported for *S. cerevisiae* using similar assays lie in the range of 1 per 10 000 000 to 1 per 100 000 000 divisions in unperturbed conditions (68), whereas studies exploiting the endogenous counter-selectable non-fragile gene *HPRT* indicate mutation rates at 6 per 10 000 000 or 8 per 1 000 000 divisions in mismatch repair proficient or deficient HCT116 colon cancer cell lines, respectively (69). Glover et al. found that certain mammalian fragile sites hold deletions in one out of three clones after a few cell divisions in the presence of APH hinting that the overall mutation rate at CFSs may be substantially higher than the rate reported here (70). Notably, the *HyTK* reporter gene used to detect deletions was inserted toward the start of the *PARK2* gene (Figure 5A). Generally, CFS deletions have a tendency to form around the middle of the large gene (53) suggesting that only a fraction of APH-induced deletions in the *PARK2* gene is picked up by our fluctuation analysis. This may explain why our rates are lower than expected. Regarding mutation rates at non-fragile sites, the assay presented here estimates the mutation rate at the OVAL control locus in the range of 4 to 15 mutations per 1 000 000 cell divisions with little change in response to replication stress. However, we failed to detect any mutation in the *HyTK* coding region in many of the ganciclovir-resistant clones suggesting that the developed assay overestimates the actual mutation rate at the *OVAL* locus. In conclusion, the mutation rate at *PARK2* is at least 10-fold higher than at a non-fragile locus.

The total size of the haploid chicken genome is 1.3 × 10^9 bp, which is roughly one third of the haploid human genome (48). Correspondingly, the largest genes in the chicken genome are just below 1 Mb, whereas the largest human genes are >2 Mb. Crocodilians, the closest living relatives of birds, have genome sizes in the range of 2–3 Gb (57). The general reduction of genome size observed in birds is thought to be an adaption to the high metabolism necessary for flight (71). The fact that bats have the smallest genomes among mammals supports this notion (71). The genome size reduction results both from elimination of transposable elements as well as a general reduction of intergenic as well as intragenic intronic sequences (72). Nevertheless, we find that introns in the extremely large fragile genes, to some extent have resisted reduction through evolution despite being inherently prone to deletion events (in somatic cells). Thus, the findings presented here have wide implications by showing that the large sizes of CFS genes are conserved between mammals and birds, indicating that the large gene structure of CFS genes has an important biological function. In this respect it is worth noting that expression of extremely large genes is a characteristic of neurons (73). Intriguingly, long neuronal genes are hotspots for DNA breaks in neuronal stem/progenitor cells, and this regional genomic instability has been suggested to play a role in neuronal diversification (54).

AVAILABILITY

FANCD2 ChIP-seq and RNA-seq data generated in this study is submitted to ArrayExpress, European Nucleotide Archive (ENA), with the respective accession numbers E-MTAB-5880 and E-MTAB-5881. To view ChIP-seq data in UCSC, bedgraph files are available. (Experiment G2/M: http://clients.galaxy.bio.ku.dk/lisby/fragile/galgal5.chp.sample43.difference.bedGraph. Exp:async: http://clients.galaxy.bio.ku.dk/lisby/fragile/galgal5.chp.sample87.difference.bedGraph).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Jette Bornholdt Lange for assistance with the RNA quality check and Olaf Nielsen for help with elutriation. Furthermore we are grateful to Christine Farr and Jakob Nilsson for sharing reagents and to Sebastian H.N. Munk and Vasileios Voutsinos for helpful discussions.

Author contributions: Conceptualization, V.H.O. Methodology, C.P., V.H.O., S.A.S., N.R.H., B.L.T., M.D., A.S.O. Validation, C.P., V.H.O. Formal analysis, S.A.S., N.R.H., C.P., V.H.O. Investigation, C.P., V.H.O., S.A.S., N.R.H., A.S.O. Data curation, N.R.H., S.A.S. Writing Original Draft, V.H.O. Writing—review and editing, V.H.O., M.L., C.P. and N.R.H. Supervision, V.H.O., M.L. Project administration, V.H.O., M.L. Funding acquisition, M.L.

FUNDING

Villum Foundation; Lundbeck Foundation (to M.L., V.H.O.); European Research Council [ERCStG, 242905 to M.L., V.H.O., C.P.); Danish National Research Foundation (DNRF115) to M.L.; French Association pour la Recherche sur le Cancer [S1220130607073 to M.D., B.L.T.]; Agence Nationale de la Recherche [13-BSV6-0008-01]; Institut National du Cancer [2013-103]. Funding for open access charge: Villum fonden.

Conflict of interest statement. None declared.

REFERENCES

1. Minocherhomji,S. and Hickson,I.D. (2014) Structure-specific endonucleases: guardians of fragile site stability. *Trends Cell Biol.*, **24**, 321–327.
2. Debatisse,M., Le Tallec,B., Letessier,A., Dutrillaux,B. and Brison,O. (2012) Common fragile sites: mechanisms of instability revisited. *Trends Genet.*, **28**, 22–32.
3. Torres-Rosell,J., De Piccoli,G., Cordon-Preciado,V., Farmer,S., Jarmuz,A., Machin,F., Pasero,P., Lisby,M., Haber,J.E. and Aragon,L. (2007) Anaphase onset before complete DNA replication with intact checkpoint responses. *Science*, **315**, 1411–1415.
4. Gallina,I., Christiansen,S.K., Pedersen,R.T., Lisby,M. and Oestergaard,V.H. (2016) TopBP1-mediated DNA processing during mitosis. *Cell Cycle*, **15**, 176–183.
5. Fragkos,M. and Naim,V. (2017) Rescue from replication stress during mitosis. *Cell Cycle*, **16**, 613–633.
6. Glover,T.W., Berger,C., Coyle,J. and Echo,B. (1984) DNA polymerase alpha inhibition by aphidicolin induces gaps and breaks at common fragile sites in human chromosomes. *Hum. Genet.*, **67**, 136–142.
7. Bignell,G.R., Greenman,C.D., Davies,H., Butler,A.P., Edkins,S., Andrews,J.M., Buck,G., Chen,L., Beare,D., Latimer,C. *et al.* (2010) Signatures of mutation and selection in the cancer genome. *Nature*, **463**, 893–898.
8. Barlow,J.H., Faryabi,R.B., Callen,E., Wong,N., Malhowski,A., Chen,H.T., Gutierrez-Cruz,G., Sun,H.W., McKinnon,P., Wright,G. *et al.* (2013) Identification of early replicating fragile sites that contribute to genome instability. *Cell*, **152**, 620–632.
9. Le Tallec,B., Millot,G.A., Blin,M.E., Brison,O., Dutrillaux,B. and Debatisse,M. (2013) Common fragile site profiling in epithelial and erythroid cells reveals that most recurrent cancer deletions lie in fragile sites hosting large genes. *Cell Rep.*, **4**, 420–428.
10. Dereli-Oz,A., Versini,G. and Halazonetis,T.D. (2011) Studies of genomic copy number changes in human cancers reveal signatures of DNA replication stress. *Mol. Oncol.*, **5**, 308–314.
11. Glover,T.W., Wilson,T.E. and Arlt,M.F. (2017) Fragile sites in cancer: more than meets the eye. *Nat. Rev. Cancer*, **17**, 489–501.
12. Chan,K.L., Palmai-Pallag,T., Ying,S. and Hickson,I.D. (2009) Replication stress induces sister-chromatid bridging at fragile site loci in mitosis. *Nat. Cell Biol.*, **11**, 753–760.
13. Naim,V. and Rosselli,F. (2009) The FANC pathway and BLM collaborate during mitosis to prevent micro-nucleation and chromosome abnormalities. *Nat. Cell Biol.*, **11**, 761–768.
14. Zhang,H. and Freudenreich,C.H. (2007) An AT-rich sequence in human common fragile site FRA16D causes fork stalling and chromosome breakage in S. cerevisiae. *Mol. Cell*, **27**, 367–379.
15. Zlotorynski,E., Rahat,A., Skaug,J., Ben-Porat,N., Ozeri,E., Hershberg,R., Levi,A., Scherer,S.W., Margalit,H. and Kerem,B. (2003) Molecular basis for expression of common and rare fragile sites. *Mol. Cell Biol.*, **23**, 7143–7151.
16. Lukusa,T. and Fryns,J.P. (2008) Human chromosome fragility. *Biochim. Biophys. Acta*, **1779**, 3–16.
17. Gao,G. and Smith,D.I. (2014) Very large common fragile site genes and their potential role in cancer development. *Cell Mol. Life Sci.*, **71**, 4601–4615.
18. Le Beau,M.M., Rassool,F.V., Neilly,M.E., Espinosa,R. 3rd, Glover,T.W., Smith,D.I. and McKeithan,T.W. (1998) Replication of a common fragile site, FRA3B, occurs late in S phase and is delayed further upon induction: implications for the mechanism of fragile site induction. *Hum. Mol. Genet.*, **7**, 755–761.
19. Letessier,A., Millot,G.A., Koundrioukoff,S., Lachages,A.M., Vogt,N., Hansen,R.S., Malfoy,B., Brison,O. and Debatisse,M. (2011) Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. *Nature*, **470**, 120–123.
20. Ozeri-Galai,E., Lebofsky,R., Rahat,A., Bester,A.C., Bensimon,A. and Kerem,B. (2011) Failure of origin activation in response to fork stalling leads to chromosomal instability at fragile sites. *Mol. Cell*, **43**, 122–131.
21. Helmrich,A., Ballarino,M. and Tora,L. (2011) Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Mol. Cell*, **44**, 966–977.
22. Helmrich,A., Stout-Weider,K., Hermann,K., Schrock,E. and Heiden,T. (2006) Common fragile sites are conserved features of human and mouse chromosomes and relate to large active genes. *Genome Res.*, **16**, 1222–1230.
23. Oestergaard,V.H. and Lisby,M. (2016) Transcription-replication conflicts at chromosomal fragile sites-consequences in M phase and beyond. *Chromosoma*, **126**, 213–222.
24. Howlett,N.G., Taniguchi,T., Durkin,S.G., D'Andrea,A.D. and Glover,T.W. (2005) The Fanconi anemia pathway is required for the DNA replication stress response and for the regulation of common fragile site stability. *Hum. Mol. Genet.*, **14**, 693–701.
25. Madireddy,A., Kosiyatrakul,S.T., Boisvert,R.A., Herrera-Moyano,E., Garcia-Rubio,M.L., Gerhardt,J., Vuono,E.A., Owen,N., Yan,Z., Olson,S. *et al.* (2016) FANCD2 facilitates replication through common fragile sites. *Mol. Cell*, **64**, 388–404.
26. Garaycoechea,J.I. and Patel,K.J. (2014) Why does the bone marrow fail in Fanconi anemia? *Blood*, **123**, 26–34.
27. Magenis,R.E., Hecht,F. and Lovrien,E.W. (1970) Heritable fragile site on chromosome 16: probable localization of haptoglobin locus in man. *Science*, **170**, 85–87.
28. Yunis,J.J. and Soreng,A.L. (1984) Constitutive fragile sites and cancer. *Science*, **226**, 1199–1204.
29. Smeets,D.F. and van de Klundert,F.A. (1990) Common fragile sites in man and three closely related primate species. *Cytogenet. Cell Genet.*, **53**, 8–14.
30. Ruiz-Herrera,A., Garcia,F., Fronicke,L., Ponsa,M., Egozcue,J., Caldes,M.G. and Stanyon,R. (2004) Conservation of aphidicolin-induced fragile sites in Papionini (Primates) species and humans. *Chromosome Res.*, **12**, 683–690.
31. Burckstummer,T., Bennett,K.L., Preradovic,A., Schutze,G., Hantschel,O., Superti-Furga,G. and Bauch,A. (2006) An efficient

tandem affinity purification procedure for interaction proteomics in mammalian cells. *Nat. Methods*, **3**, 1013–1019.

32. Arakawa,H., Lodygin,D. and Buerstedde,J.M. (2001) Mutant loxP vectors for selectable marker recycle and conditional knock-outs. *BMC Biotechnol.*, **1**, 7.
33. Lupton,S.D., Brunton,L.L., Kalberg,V.A. and Overell,R.W. (1991) Dominant positive and negative selection using a hygromycin phosphotransferase-thymidine kinase fusion gene. *Mol. Cell. Biol.*, **11**, 3374–3378.
34. Gillespie,D.A. and Henriques,C. (2006) Centrifugal elutriation as a means of cell cycle phase separation and synchronisation. *Subcell. Biochem.*, **40**, 359–361.
35. Smith,K.A., Gorman,P.A., Stark,M.B., Groves,R.P. and Stark,G.R. (1990) Distinctive chromosomal structures are formed very early in the amplification of CAD genes in Syrian hamster cells. *Cell*, **63**, 1219–1227.
36. El Achkar,E., Gerbault-Seureau,M., Muleris,M., Dutrillaux,B. and Debatisse,M. (2005) Premature condensation induces breaks at the interface of early and late replicating chromosome bands bearing common fragile sites. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 18069–18074.
37. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
38. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
39. Krzywinski,M., Schein,J., Birol,I., Connors,J., Gascoyne,R., Horsman,D., Jones,S.J. and Marra,M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.
40. Nakato,R., Itoh,T. and Shirahige,K. (2013) DROMPA: easy-to-handle peak calling and visualization software for the computational analysis and validation of ChIP-seq data. *Genes Cells*, **18**, 589–601.
41. Shang,W.H., Hori,T., Martins,N.M., Toyoda,A., Misu,S., Monma,N., Hiratani,I., Maeshima,K., Ikeo,K., Fujiyama,A. *et al.* (2013) Chromosome engineering allows the efficient isolation of vertebrate neocentromeres. *Dev. Cell*, **24**, 635–648.
42. Speir,M.L., Zweig,A.S., Rosenbloom,K.R., Raney,B.J., Paten,B., Nejad,P., Lee,B.T., Learned,K., Karolchik,D., Hinrichs,A.S. *et al.* (2016) The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.*, **44**, D717–D725.
43. Buerstedde,J.M. (2006) Subcloning Dt40 by limiting dilution. *Subcell Biochem.*, **40**, 393–394.
44. Hall,B.M., Ma,C.X., Liang,P. and Singh,K.K. (2009) Fluctuation analysis CalculatOR: a web tool for the determination of mutation rate using Luria-Delbruck fluctuation analysis. *Bioinformatics*, **25**, 1564–1565.
45. Luria,S.E. and Delbruck,M. (1943) Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, **28**, 491–511.
46. Huerta-Cepas,J., Szklarczyk,D., Forslund,K., Cook,H., Heller,D., Walter,M.C., Rattei,T., Mende,D.R., Sunagawa,S., Kuhn,M. *et al.* (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.*, **44**, D286–D293.
47. Huerta-Cepas,J., Forslund,K., Pedro Coelho,L., Szklarczyk,D., Juhl Jensen,L., von Mering,C. and Bork,P. (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.*, **34**, 2115–2122.
48. Wallis,J.W., Aerts,J., Groenen,M.A., Crooijmans,R.P., Layman,D., Graves,T.A., Scheer,D.E., Kremitzki,C., Fedele,M.J., Mudd,N.K. *et al.* (2004) A physical map of the chicken genome. *Nature*, **432**, 761–764.
49. Pedersen,R.T., Kruse,T., Nilsson,J., Oestergaard,V.H. and Lisby,M. (2015) TopBP1 is required at mitosis to reduce transmission of DNA damage to G1 daughter cells. *J. Cell Biol.*, **210**, 565–582.
50. Bednarek,A.K., Laflin,K.J., Daniel,R.L., Liao,Q., Hawkins,K.A. and Aldaz,C.M. (2000) WWOX, a novel WW domain-containing protein mapping to human chromosome 16q23.3-24.1, a region frequently affected in breast cancer. *Cancer Res.*, **60**, 2140–2145.
51. Metsalu,T. and Vilo,J. (2015) ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Res.*, **43**, W566–W570.
52. Kantidakis,T., Saponaro,M., Mitter,R., Horswell,S., Kranz,A., Boeing,S., Aygun,O., Kelly,G.P., Matthews,N., Stewart,A. *et al.* (2016) Mutation of cancer driver MLL2 results in transcription stress and genome instability. *Genes Dev.*, **30**, 408–420.
53. Wilson,T.E., Arlt,M.F., Park,S.H., Rajendran,S., Paulsen,M., Ljungman,M. and Glover,T.W. (2015) Large transcription units unify copy number variants and common fragile sites arising under replication stress. *Genome Res.*, **25**, 189–200.
54. Wei,P.C., Chang,A.N., Kao,J., Du,Z., Meyers,R.M., Alt,F.W. and Schwer,B. (2016) Long neural genes harbor recurrent DNA break clusters in neural stem/progenitor cells. *Cell*, **164**, 644–655.
55. Wurtele,H., Kaiser,G.S., Bacal,J., St-Hilaire,E., Lee,E.H., Tsao,S., Dorn,J., Maddox,P., Lisby,M., Pasero,P. *et al.* (2012) Histone h3 lysine 56 acetylation and the response to DNA replication fork damage. *Mol. Cell Biol.*, **32**, 154–172.
56. Buerstedde,J.M. and Takeda,S. (1991) Increased ratio of targeted to random integration after transfection of chicken B cell lines. *Cell*, **67**, 179–188.
57. Green,R.E., Braun,E.L., Armstrong,J., Earl,D., Nguyen,N., Hickey,G., Vandewege,M.W., St John,J.A., Capella-Gutierrez,S., Castoe,T.A. *et al.* (2014) Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science*, **346**, 1254449.
58. Sarni,D. and Kerem,B. (2016) The complex nature of fragile site plasticity and its importance in cancer. *Curr. Opin. Cell Biol.*, **40**, 131–136.
59. Martinez-Balbas,M.A., Dey,A., Rabindran,S.K., Ozato,K. and Wu,C. (1995) Displacement of sequence-specific transcription factors from mitotic chromatin. *Cell*, **83**, 29–38.
60. Core,L.J., Waterfall,J.J. and Lis,J.T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845–1848.
61. Paulsen,M.T., Veloso,A., Prasad,J., Bedi,K., Ljungman,E.A., Tsan,Y.C., Chang,C.W., Tarrier,B., Washburn,J.G., Lyons,R. *et al.* (2013) Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced proinflammatory response. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 2240–2245.
62. Li,X. and Manley,J.L. (2005) Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. *Cell*, **122**, 365–378.
63. Prado,F. and Aguilera,A. (2005) Impairment of replication fork progression mediates RNA polII transcription-associated recombination. *EMBO J.*, **24**, 1267–1276.
64. Costantino,L. and Koshland,D. (2015) The Yin and Yang of R-loop biology. *Curr. Opin. Cell Biol.*, **34**, 39–45.
65. Snyder,M., Sapolsky,R.J. and Davis,R.W. (1988) Transcription interferes with elements important for chromosome maintenance in Saccharomyces cerevisiae. *Mol. Cell. Biol.*, **8**, 2184–2194.
66. Looke,M., Reimand,J., Sedman,T., Sedman,J., Jarvinen,L., Varv,S., Peil,K., Kristjuhan,K., Vilo,J. and Kristjuhan,A. (2010) Relicensing of transcriptionally inactivated replication origins in budding yeast. *J. Biol. Chem.*, **285**, 40004–40011.
67. Gros,J., Kumar,C., Lynch,G., Yadav,T., Whitehouse,I. and Remus,D. (2015) Post-licensing specification of eukaryotic replication origins by facilitated Mcm2-7 sliding along DNA. *Mol. Cell*, **60**, 797–807.
68. Lang,G.I. and Murray,A.W. (2011) Mutation rates across budding yeast chromosome VI are correlated with replication timing. *Genome Biol. Evol.*, **3**, 799–811.
69. Glaab,W.E. and Tindall,K.R. (1997) Mutation rate at the hprt locus in human cancer cell lines with specific mismatch repair-gene defects. *Carcinogenesis*, **18**, 1–8.
70. Arlt,M.F., Ozdemir,A.C., Birkeland,S.R., Wilson,T.E. and Glover,T.W. (2011) Hydroxyurea induces de novo copy number variants in human cells. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 17360–17365.
71. Wright,N.A., Gregory,T.R. and Witt,C.C. (2014) Metabolic 'engines' of flight drive genome size reduction in birds. *Proc. Biol. Sci.*, **281**, 20132780.
72. Zhang,G., Li,C., Li,Q., Li,B., Larkin,D.M., Lee,C., Storz,J.F., Antunes,A., Greenwold,M.J., Meredith,R.W. *et al.* (2014) Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, **346**, 1311–1320.
73. Gabel,H.W., Kinde,B., Stroud,H., Gilbert,C.S., Harmin,D.A., Kastan,N.R., Hemberg,M., Ebert,D.H. and Greenberg,M.E. (2015) Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature*, **522**, 89–93.