Transcription-Replication Conflict Shapes DNA Break Dynamics

Vivien Ionasz[1,2,3,8], Lorenzo Corazzi[1,2,3,8], Sergej Andrejev[1,2], Athanasios Vouzas[4,6], Li-Chin Wang [1,2], Marco Giaisi[1,2], Giulia Di Muzio[1,2,3], Boyu Ding[1,2,5], Jinrong Fu[1,2,5], Amina Abdelbaki[1,2,3], David M. Gilbert [6], and Pei-Chi Wei[1,2,3,7,9]

1. Brain Mosaicism and Tumorigenesis, German Cancer Research Center, 69120 Heidelberg, Germany.
2. Structural and Functional Genomics, German Cancer Research Center, 69120 Heidelberg, Germany.
3. Faculty of Bioscience, Ruprecht Karl University of Heidelberg, 69120 Heidelberg, Germany.
4. Department of Biological Science, Florida State University, FL 32306, United States of America.
5. Faculty of Medicine, Ruprecht Karl University of Heidelberg, 69120 Heidelberg, Germany.
6. San Diego Biomedical Research Institute, San Diego, CA 92121, United States of America.
7. Interdisciplinary Center for Neurosciences, Ruprecht Karl University of Heidelberg, 69120 Heidelberg, Germany.
8. These authors contributed equally to the manuscript.
9. To whom correspondence should be addressed: Tel: +49(0)6221423251; email: p.wei@dkfz-heidelberg.de.

Abstract

Double strand breaks (DSBs) originating from transcription and replication conflict (TRC) sites are prone to rearrangements. Through the application of a capture-ligation assay on mouse neural progenitor cells experiencing replication stress, we unveiled that interactions between transcription and replication fork architecture dictate DSB location and orientation. Specifically, telomere-to-centromere forks generate telomere-connected DSBs, while centromere-to-telomere forks lead to centromere-connected DSBs in genomes that replicate during median and late stages. This pattern, however, reverses in early-replicating DNA. Mapping nascent RNA and RNA polymerase activity revealed that head-to-head interactions between replication and transcription machineries elevate DSBs by 30% compared to co-directional interactions. By deleting promoter elements in the moderately transcribed *Ctnna2* and *Nrxn1* genes, we demonstrated that the removal of transcription suppressed DSB. However, titration of expression at the *Ptn* locus demonstrated that transcription level alone was insufficient to induce DSB. Our findings underscore how transcription and replication directionality impact genome structural stability.

INTRODUCTION

DNA double strand breaks (DSBs) play a vital role in cellular physiology. During replication stress, they emerge at stalled forks to reconfigure and restart the process[1]. Disrupting this leads to slower DNA replication and genome instability[2]. In mice, this instability can cause tumor formation[3]. As replication stress intensifies, excessive DSBs at forks became sensitive for rearrangements[4]. Structure-specific DNA nucleases at stalled forks recognize Y-shaped DNA junctions and make cuts[5], yielding a single DNA end in biochemical tests[6–9]. According to the proposed model, DSBs at the replication fork are of a single-ended nature, implying that each nuclease activity produces only a solitary DSB end at the Y-shaped DNA junction. These DSB ends possess inherent orientations, attaching themselves to either centromeric or telomeric sequences on mammalian chromosomes. Gaining an understanding of these orientations holds the potential to illuminate the processes of genome rearrangements under conditions of replication stress. However, the exact orientation of DSB ends at the replication fork within living cells remains an unanswered question.

The challenge of fork stalling intensifies when transcription and replication machineries operate in the same genomic regions. In mammalian cells, long genes have been identified as sites prone to conflicts between transcription and replication, known as transcription-replication conflicts (TRC)[10–12]. Within these TRC zones, DNA breaks are induced by specialized DNA endonucleases, resulting in the conversion of TRC sites into common fragile sites (CFS) [13,14]. CFSs are associated with copy number losses in cancer genomes, although the precise mechanism remains incompletely understood. One proposed explanation suggests that CFSs could be situated within "double fork failure" zones[11,15], where two inward-moving replication forks traverse the entire gene, leading to DNA replication termination midway through the gene during the S phase. In the context of this "double fork failure" model, it has been hypothesized that repair pathways relying on sequence homology might contribute to copy number loss[16]. However, it's important to note that repair pathways based on sequence homology do not exclusively result in deletions. While it was suggested that the orientation of DNA ends might

play a role in copy number loss at CFS [6,15], additional research is required to ascertain the degree to which DSB ends align to promote deletions.

Furthermore, within highly transcribed genomes, there exists the phenomenon where the DNA template and RNA transcripts occasionally hybridize, giving rise to a structure known as an R-loop [17]. The accumulation of R-loops has been linked to DNA instability[18]. Notably, the impact of R-loops on DNA breakage becomes more pronounced in instances of "head-on" conflicts—situations where transcription and replication machineries operate in opposing directions. Experimental models utilizing episomes have proposed that head-on collisions enhance R-loop accumulation, thereby amplifying DNA breaks[18]. However, considering that only a relatively small portion of the transcribed mammalian genome, approximately 2-15%, is susceptible to R-loop buildup[19], an inquiry arises as to whether the densities of DSBs at other TRC sites exhibit similar kinetics in the head-on direction. Furthermore, the question remains whether transcription activity alone is capable of initiating DSB formation at TRC sites, warranting further investigation.

Our investigation focused on DSB ends within a recently identified class of TRC, namely the recurrent DNA break clusters (RDCs). RDC was initially observed in mouse neural progenitor cells subjected to low-dose aphidicolin (APH), a DNA polymerase inhibitor[20,21]. RDCs distinguish themselves from CFS by exhibiting distinct replication characteristics. While 62% of RDCs occur in regions that replicate at the late S phase, 38% are associated with earlier replication [20,21]. Significantly, RDCs are not confined solely to neural progenitor cells; they have also been identified in other cell types, including mouse embryonic stem (ES) cells [22], as well as in vitro-derived human oligodendrocyte progenitors and astrocyte progenitors[23]. Although the specific genomic positions of RDCs may vary between different cell types, a consistent pattern emerges—RDCs consistently localize to actively transcribed genes. These findings suggest the presence of a shared mechanism driving RDC formation across mammalian cells. Notably, comprehensive mapping of RDCs has been achieved across all

chromosomes in mouse neural progenitor cells, whereas in other cells, only the most prominent RDCs have been successfully mapped[21,22]. To ensure a thorough exploration of RDCs, we opted to conduct our subsequent experiments using mouse neural progenitor cells.

In our study, we have revealed a distinctive pattern of DSB end orientation at RDCs. Replication forks moving toward the centromere generated telomeric-oriented DSB ends, while those heading toward telomeres produced centromeric-oriented DSB ends, aligning with biochemical predictions. We found that the orientation of DSB ends outward moving fork directions were opposite the proposed model. This orientation corresponded to DNA replication fork slowing and highlights the link between DSB end density and replication dynamics. Additionally, we found that head-to-head interactions between replication and transcription machinery increased DSB formation by 30% compared to co-directional interactions. Moreover, our systematic exploration emphasized the role of transcription in DSB formation at RDCs, illuminating how the interplay of transcription and replication directions impacts genome stability. In summary, our study provides insights into the intricate mechanisms of DSB formation at RDCs, enhancing the understanding of genome stability maintenance.

RESULTS

Replication of the DNA Sequences Underneath RDCs Involves Dynamic Fork Directions.

To elucidate the DNA replication directionality within DNA sequences underlying recurrent DNA break clusters (RDCs), we established connections between initiation and termination zones. The determination of initiation zones employed an algorithm developed for this purpose[24], using high-resolution replication sequencing (Repli-seq) data derived from mouse ES cell-derived NPCs[25]. Notably, termination zones exhibited a higher degree of dynamism compared to initiation zones. To address this, we designed a convolutional neural network to aid in characterizing termination zones (Fig. 1A and Methods). The network was

trained using diverse features extracted from published mouse ES cell Okazaki-sequencing (OK-seq) datasets, as well as high-resolution Repli-seq datasets[25,26]. Post-training, our model achieved an impressive 80% recall rate with precision exceeding 80% within a 50kb window size (Fig. 1B). We then applied the trained convolutional neural network model to high-resolution Repli-seq data from mouse NPCs[25]. This application led to the definition of 2922 termination zones. Subsequently, through the connection of initiation and termination zones, we identified 5844 replication fork directions in NPCs, each with a median length of 375 kb (Fig. 1C). When mapping these replication fork directions to the mouse genome, we discovered that 48% of RDCs comprised two inward-moving forks, while 23% of RDCs were situated within unidirectional replication fork territories. Moreover, 7.9 % of RDCs resided in regions characterized by outward-moving forks, and 17.1% of RDCs exhibited complex fork compositions. Additionally, for six RDCs located on the X chromosome, replication direction could not be determined due to limited Repli-seq data depth. Intriguingly, 7.4% of RDCs coincided with genomic regions displaying biphasic replication patterns (see Table S1).

Oriented Single-Ended DSBs at RDCs Correspond to Replication Fork Directions

We hypothesize that DSB ends within RDCs exhibit specific orientations aligned with replication fork directions. We utilized linear-amplification mediated, high-throughput genome-wide translocation sequencing (LAM-HTGTS)[27] to determine these orientations. This approach involved a designated "bait" DSB end induced by sequence-specific enzymes like CRISPR/Cas9 ([28] and Fig. S1A), revealing translocations between the viewpoint and other "prey" DSB ends. Importantly, the orientation of these "prey" DSBs was retained through linear-amplification-mediated PCR (Fig. S1A). DSB ends associated with centromeres aggregate to form centromeric DSB (Dcen) peaks, while DSB ends attaching to telomeres result in telomeric DSB (Dtel) peaks. When the peaks of Dcen and Dtel align precisely, it indicates that the interval between Dcen and Dtel is extremely brief, spanning only a few base pairs (Fig. S1B). On the other hand, if the two peaks are separate, it signifies that Dcen and Dtel ends are located individually on opposite sides on the genome, with a significant interval existing between them

(Fig. S1C). In cases where only either Dcen or Dtel is found within the specific genomic region, one would observe a distinctive single-peak pattern without overlap, indicating the orientation of the ends (Fig. S1D). For efficient recovery of genome-wide "prey" DSBs at RDCs, we selected Xrcc4 and Tp53-deficient mouse ES-cell-derived NPCs [20], with Tp53 knockout preventing NPC death caused by Xrcc4 deficiency[20]. In our annotation process, we encompassed all DSBs within a two-megabase range from the CRISPR/Cas9-induced viewpoint and those distributed across other chromosomes. Notably, analysis revealed an uneven distribution of centromeric and telomeric DNA ends on the viewpoint-containing chromosome, while other chromosomes maintained balanced distributions. Consequently, DSBs on the viewpoint-containing chromosomes were excluded from DSB end orientation analyses.

We proposed that fork stalling at inward-moving forks yields centromeric DSB ends at right-moving forks and telomeric DNA ends at left-moving forks (Fig. 2A). To test this, we initially analyzed DSB distribution at *Ctnna2* and *Prkg1*, both RDC-containing gene loci replicated by inward-moving forks. At *Ctnna2*, Dcen and Dtel peaks aligned with right-moving and left-moving replication fork directions, respectively (Fig. 2B). A similar "twin-peak" pattern was observed at the *Prkg1* locus (Fig. 2B) and 53 other inward-moving fork RDCs (Table S1). These "twin-peak" RDCs exhibited a Dcen-to-Dtel distance of 100 kb to 1 Mb (median 280 kb, Fig. S2A).

We then investigated RDCs situated within DNA sequences governed by unidirectional forks (Fig. 2C). Our hypothesis proposed that RDCs in these loci would exhibit enriched DSB end orientations matching the replication fork direction. To validate this, we scrutinized DSB end distributions at *Magi1* and *Large*, both RDC-containing genes controlled by right-moving forks (Fig. 2D). As anticipated, Dcen density, but not Dtel, displayed enrichment in both gene loci. Out of the 35 RDCs within unidirectional fork zones, 20 exhibited the "single-peak" signature, where DNA end density aligned with the corresponding fork direction in that region.

Next, we turned our focus to genomic regions governed by outward-moving forks (Fig. 2E). These regions are replicating early, akin to early replicating fragile sites [29]. Within these zones, the distributions of Dcen and Dtel at RDCs exhibited overlapping patterns, as shown by the example of an RDC on chromosome 10 (Fig. 2F, left), or presented as twin peaks with gap intervals less than 100 kb (Fig. 2F, right). At RDCs that expand outward, the Dcen peak correlated with a fork moving to the left, while the Dtel peak corresponded to a fork moving to the right. This observation indicated that the DSB ends within this region attached to genomic sequences that were still pending replication.

Moving forward, we explored *Auts2* and *Sdk1* loci, where genomic DNA displayed biphasic replication (Fig. 2G, H). Although the neural network could only determine the inward-moving fork directions, high-resolution Repli-seq suggested pairs of inward and outward moving forks at these loci. Interestingly, a "double twin-peak" RDC pattern emerged at *Auts2* and *Sdk1*, with two Dcen/Dtel peaks each. Notably, the outer Dcen/Dtel peaks corresponded to inward-moving fork landscapes, while the inner peaks corresponded to outward-moving forks. Among the 9 RDCs replicated by two inward-moving forks, 5 exhibited the "double twin-peak" pattern (Table S1). The other RDCs distribution of Dcen and Dtel varied from one locus to another (Table S1). Collectively, our findings revealed a broad variability in the distribution of oriented DSB ends within RDCs.

We conclude that the distribution of Dcen and Dtel correlates with fork direction, with the highest correlation observed in regions governed by inward-moving and unidirectional forks.

Position and Density of the Oriented DSB Ends Corresponded to DNA Replication

Previous studies have demonstrated that ATR inhibition with VE-821 induces a global slowdown of replication forks[30]. In cells that have already been subjected to aphidicolin treatment, inhibiting ATR would add additional stress to the replication forks. Consequently, at loci with inward-moving replication forks, ATR inhibition would result in an increased gap between right-moving and left-moving forks. This effect was evident at the *Ctnna2* locus, where

the separation between Dcen and Dtel peaks expanded from 160 kb to over 500 kb (Fig. 3A). Hence, ATR-mediated fork slowing affected the position of DSB ends in RDCs governed by inward-moving forks. Conversely, ATR inhibition has been shown to promote dormant origins firing around active replication forks[31]. It is not clear whether early or late replicating DNA react the same to ATR inhibition. To test this, we examined oriented DSB end density at the "single-peak" loci, *Magi1* and *Large*. *Magi1* underwent early replication, while *Large* replicated in the middle of S phase. Under ATR inhibition, a second Dcen peak emerged at the *Magi1* locus (Fig. 3B), suggesting that additional origins within the *Magi1* gene locus were activated. Thus, the newly initiated right-moving forks stalled, resulting in an additional Dcen peak. Conversely, no additional Dcen peak was observed at the *Large* locus (Fig. 3C), indicating that ATR inhibition did not affect the replication fork beneath *Large*. we have demonstrated that genomic loci undergoing early replication in the S phase, as opposed to regions replicating during the median or late phases, display dormant origin activation upon ATR inhibition.

Next, we investigated whether the density of single-ended DSBs correlates positively with the level of replication stress. We analyzed RDCs in NPCs treated with varying concentrations of aphidicolin (0.2, 0.3, 0.4, and 0.6 μM). In the case of the *Ctnna2* RDC, DSB density increased with rising aphidicolin dosage, from 45.8 DSB per thousand per megabase (DTM) in NPCs treated with 0.3 μM aphidicolin to 165.2 DTM in NPCs treated with 0.6 μM aphidicolin (Fig. 3D). Subsequently, we evaluated 137 RDCs containing a minimum of five DSBs (Fig. 3E), observing a similar trend of aphidicolin dosage-dependent DSB density elevation across the NPC genome. Overall, our findings strongly suggest a correlation between the density of oriented DSB ends at RDCs and the slowdown of DNA replication forks.

Enhanced DSB Densities in RDCs Due to Head-On Collisions.

We then delved into the prevalence of DSB ends resulting from head-on versus codirectional collisions at RDCs. For this analysis, we utilized global run-on sequencing (GRO-seq) to determine transcription strandness[32]. At the *Ctnna2* locus, nascent RNA sequences aligned with the Crick strand, indicating a left-moving transcription direction. This led to head-

on conflict at the right-moving fork and codirectional collision at the right-moving transcription site. Notably, in the head-on region of the *Ctnna2* locus, Dcen density significantly increased from 26 to 99.3 DTM in NPCs treated with 0.3 μM or 0.6 μM aphidicolin. In comparison, the codirectional counterpart displayed a more moderate rise in Dtel density, increasing from 20 DTM to 65.9 DTM under the same conditions (density calculated from data shown in Fig. 3D). Similarly, head-on conflicts contributed to higher DSB densities at another RDC-containing gene locus, *Tenm3*. In NPCs, *Tenm3* has two actively transcribed isoforms in a left-moving direction (Fig. 4A). The long isoform experienced head-on collision with the right-moving fork and co-directional conflict with the left-moving forks. In contrast, the short *Tenm3* isoform only encountered head-on conflict. The head-on DSB density increased from 7.6 DTM in NPCs without aphidicolin treatment to 25.5 DTM with 0.6 μM aphidicolin treatment. In the co-directional counterpart, the increase in RDC density was more moderate, rising from 7.0 DTM in DSMO-treated cells to 10.2 DTM with 0.6 μM aphidicolin treatment (Fig. 4A).

To comprehensively evaluate head-on and codirectional contributions at a genome-wide scale, we acknowledged that not all RDCs feature two long convergent forks (Table S1). This implies that some RDCs are influenced primarily by head-on or co-directional collisions. To address this diversity, we devised a strategy to assess RDCs of various configurations. We quantified the proportional changes in Dcen and Dtel within the head-on and co-directional regions (Fig. 4B). The ideal scenario would yield a 50% DSB proportion for both head-on and co-directional contributions. Deviations from this 50% mark indicate differential contributions. Our analysis encompassed 85 RDCs with more than seven DSBs, revealing an increasing DSB contribution from head-on colliding replication forks in a dose-dependent manner upon aphidicolin treatment (proportion of head-on, P_{HO}, from 50% to around 70%, Fig. 4C). Conversely, at co-directional replication forks, the DSB proportion (proportion of codirectional, P_{CD}) decreased in low-dose aphidicolin-treated NPCs, declining from 50% in DMSO samples to approximately 40% in aphidicolin-treated samples (Fig. 4C), with co-directional DSB contribution remaining consistently low at the same proportion. Notably, the most pronounced

disparity between head-on and codirectional contributions was achieved at 30% under the 0.6 μM aphidicolin condition. In summary, our findings robustly demonstrate that head-on conflict significantly elevates DSB density at RDCs.

A recent study has highlighted that RDCs in primary isolated neural stem and progenitor cells are largely devoid of R-loop accumulation [33]. To verify this observation, we employed the DNA:RNA immunoprecipitation (DRIP) sequencing (DRIP-seq) assay in ES cell-derived NPCs. Our findings confirmed minimal R-loop accumulation in RDCs, with only five RDCs exhibiting more than 10 DRIP-seq peaks, indicative of R-loop accumulation (Table S1). Among the RDCs analyzed, 53 contained 1-9 DRIP-seq peaks, while 93 RDCs showed no detectable DRIP-seq peaks (Table S1). Notably, R-loop presence was absent in both the *Ctnna2* and *Tenm3* gene loci, underscoring that the observed head-on bias remained independent of R-loop accumulation. Collectively, our results indicate that R-loop accumulation exerts minimal to no discernible effect on the pronounced global head-on bias observed in RDCs.

Critical Role of Prolonged Transcription in *Ctnna2* Locus RDC Formation

Given the active transcription within all RDCs, we postulated that transcription is pivotal for RDC formation. To investigate this, we harnessed CRISPR/Cas9-mediated genome editing to disrupt transcriptional activity at the *Ctnna2* locus. Initially, we eliminated the entire *b* allele in Xrcc4/p53-deficient mouse ES cells, creating Ctnna2-a/Δ ES cells (depicted in Fig. 5A). Subsequently, we deleted *cis* regulatory elements and the region proximal to the transcription start site (TSS) of *Ctnna2* on the *a* allele, generating Ctnna2-ape/Δ ES cells (Fig. 5A). Validation of these deletions was confirmed via Sanger sequencing (Fig. S3 A, B), and allelic ratios of the *Ctnna2* gene body and its regulatory element were quantified using TaqMan qPCR (Fig. S3C). Subsequent, parental, Ctnna2-a/Δ, and Ctnna2-ape/Δ mouse ES cells were *in vitro* differentiated to NPCs. In the *Ctnna2* locus of Ctnna2-a/Δ NPCs, substantial nascent

transcripts were detected, while they were absent in Ctnna2-ape/Δ NPCs (Fig. 5B). Because nascent transcript is corresponding to transcription activity, we concluded that *Ctnna2* was not actively transcribed in the Ctnna2-ape/Δ NPCs. We then assessed DSB density at the *Ctnna2* locus across the parental, Ctnna2-a/Δ, and Ctnna2-ape/Δ NPCs. To enhance DSB recovery, we employed a LAM-HTGTS viewpoint positioned just seven Mb away from the *Ctnna2* locus on mouse chromosome 6. As anticipated, the number of DSBs in Ctnna2-a/Δ NPCs decreased by half compared to parental NPCs (Fig. 5C). More strikingly, RDCs were entirely absent at the *Ctnna2* locus in Ctnna2-ape/Δ NPCs. This contrasted with another RDC-containing gene, *Foxp1*, where DSB reduction was not observed (Fig. 5D). This locus-specific loss of RDC at the *Ctnna2* gene underscores that transcription activity is a pivotal determinant for RDC formation within the *Ctnna2* locus.

Essential Role of Full-Length Transcription in RDC Formation at the *Nrxn1* Locus

We explored the pivotal contribution of transcription to the formation within the *Neurexin 1* (*Nrxn1*) gene locus. Notably, Nrxn1 encompassed two distinct transcript isoforms, spanning 1.2 Mb and 422 kb, respectively, each under the regulation of discrete promoters. To elucidate the impact of transcription elongation on RDCs, we employed targeted genome editing to eliminate the entire *b* allele of *Nrxn1*, generating Nrxn1-a/Δ ES cells. Additionally, we selectively removed regulatory elements housing the promoter for the long Nrxn1 isoform on the *a* allele, resulting in Nrxn1-ape/Δ ES cells (Fig. 6A). The fidelity of allele-specific deletions was confirmed using previously described methods (Fig. S4 A-C). Nascent RNA originating from the long Nrxn1 transcripts remained detectable in Nrxn1-a/Δ NPCs (Fig. 6B). Conversely, Nrxn1-ape/Δ NPCs exhibited nascent transcripts exclusively corresponding to the short Nrxn1 isoform. By employing the LAM-HTGTS approach on the chromosome housing *Nrxn1*, we assessed the DSBs within RDCs in response to mild aphidicolin treatment. Strikingly, RDCs persisted in Nrxn1-a/Δ NPCs, while they were entirely absent in Nrxn1-ape/Δ NPCs (Fig. 6B,C). The site-specific nature of this effect was emphasized by the unaffected status of RDC in *Ptprm*

gene locus (Fig. 6D). Our analysis substantiates that comprehensive transcriptional activity is indispensable for RDC formation at the *Nrxn1* locus.

Transcriptional inhibition has been correlated with shifts in replication timing from late to early at CFS[34]. To investigate whether transcriptional inhibition similarly impacted replication timing at the *Ctnna2* and *Nrxn1* loci, we conducted two-fraction replication sequencing to determine replication timing. Notably, transcriptional inhibition minimally advanced replication timing at the *Ctnna2* locus (Fig. S4D). Furthermore, the replication timing at the *Nrxn1* locus remained largely unaltered between Nrxn1-a/Δ and Nrxn1-ape/Δ NPCs (Fig. S4E, with p values provided). Thus, the absence of RDC formation was not attributed to a shift in replication timing from late to early.

Active Transcription Failed to Induce An RDC at *Ptn*

Transcription of the short *Nrxn1* isoform was inadequate in inducing RDC formation at the *Nrxn1* locus (Fig. 6). This discrepancy led us to postulate that transcriptional activity alone might not be adequate for RDC induction. To rigorously assess this hypothesis, we employed a gain-of-function approach to activate transcription of the *Ptn* gene in mouse ES cells. *Ptn* was chosen due to its intriguing behavior during the ES cell-to-NPC transition, where it undergoes concurrent shifts in replication timing and transcription status[35]. Notably, *Ptn* also represents a robust RDC-containing gene in ES cell-derived NPCs (Fig. S5A), in contrast to its silent transcription and RDC absence in mouse ES cells before replication stress[22]. To enable *Ptn* expression in mouse ES cells, we replaced the native *Ptn* promoter with either the human PGK (hPGK) or mouse PGK (mPGK) promoters on the F121 allele of F121-9/CAST mouse ES cell lines (Fig. 7A). Successful Ptn transcription was verified through RT-qPCR (Fig. S5B), with Ptn expression levels in mPGK and hPGK lines approximately 40-fold higher than in the endogenous *Ptn* promoter line (Fig. S5B). Furthermore, GRO-seq analyses confirmed the active transcription of the full *Ptn* gene length (Fig. 7B). To capture DNA breaks at maximum efficiency, we used a LAM-HTGTS viewpoint at the chromosome where *Ptn* located. Interestingly, the DNA break density at the *Ptn* locus remains at the same level in ES cells

harboring the endogenous *Ptn* promoter, mPGK or hPGK (Fig. 7C). To see if the level of transcription plays a role in RDC induction, we further compared the DNA break density in *Ptn* controlled by a tetracycline responding element (TRE) promoter. This promoter activates transcription in a doxycycline-dosage-dependent manner (Fig. 7D). Despite significant increases in *Ptn* transcripts with higher doxycycline concentrations (about 120-fold compared to mPGK or hPGK-dependent transcripts), no alteration in DNA break density at the *Ptn* locus was observed (Fig. 7E). In contrast, the *Ctnna2* gene locus, located on chromosome 6 exhibited distinct DNA break clusters upon aphidicolin treatment (Fig. 7F). Our findings lead us to conclude that transcriptional activation alone in F121-9/CAST mouse ES cells is insufficient to trigger RDC induction at the *Ptn* locus.

DISCUSSION

Our findings present a groundbreaking revelation – DNA breaks within TRCs exhibit distinct orientations, aligning with replication fork directions and fork slowing patterns. This crucial observation strongly suggests that these DNA breaks occur precisely at the replication fork. Notably, in cases where RDCs are flanked by two inward-moving forks, we noted a remarkable phenomenon – DSB ends at the right-moving fork align with those at the left-moving fork (Fig. 2). We propose that this head-to-head DSB orientation within *cis* configurations may foster mechanisms reliant on homology or end joining, potentially leading to deletions. For instance, at the *Ctnna2* locus, a 160 kb intragenic interval exists between Dcen and Dtel peaks. The fusion of one Dcen and one Dtel within the *Ctnna2* gene could result in a critical exon deletion, a genomic alteration strongly associated with conditions like autistic spectrum disorder and schizophrenia in humans[36]. Such deletions could potentially emerge at a somatic level during replication stress in NPCs. Given that a significant proportion of RDC-containing genes govern cell adhesion or synaptogenesis[20,21], deletions in NPCs could reverberate to daughter neurons, thereby impacting neuronal function. In totality, our study firmly establishes the role of DNA replication direction in shaping genome rearrangements within NPCs under replication stress.

The pronounced presence of DSBs at the fork within RDCs strongly implicates the involvement of structure-specific nucleases like MUS81 or XPF in their generation[7]. MUS81 exhibits heightened activity during mid-S phase and G2/M phase, crucially aiding in chromosome segregation[37]. MUS81 is also indispensable for CFS and mitotic DNA synthesis[13,38]. We proposed that MUS81 are responsible for 62% RDC-containing genes that were replicated at mid-to late S phase. In contrast, the remaining 38% of RDC-containing genes replicate at distinct phases, where we posit that the early-S phase activation of XPF drives DSB end formation[39]. However, it's important to differentiate these late-replicating RDCs from mitotic DNA synthesis, which necessitates mechanisms like break-induced replication and translesion synthesis, reliant on 2-100 bp sequence homologies between donor and recipient strands[40–42]. While these pathways are anticipated for RDC resolution, intriguingly, over 60% of RDC DSBs, joined with the LAM-HTGTS viewpoint DSB, showcased 0-1 base microhomology[20], a profile synonymous with non-homologous end joining (NHEJ)[43]. The extent to which RDC DSBs exclusively favor NHEJ or are subject to potential repair during the ensuing G1 phase, when NHEJ predominantly acts upon DSB ends, warrants further exploration.

Our investigation revealed the occurrence of both head-on and co-directional conflicts within RDC loci during DNA polymerase slowing, mirroring observations reminiscent of R-loop accumulation in certain DNA sequences[18]. Interestingly, despite the scarcity of R-loops detected in most RDC loci, we propose that the head-on bias might be attributed to transient R-loops or other DNA-RNA hybrids. An example of such hybrids includes co-transcriptional R-loops formed between the nascent transcript and template DNA, which are of a transient nature and resolve as the nascent transcript reanneals to the template DNA[17]. Similarly, the abundant RNA-DNA hybrids formed by short RNA primers during DNA replication, notably on replication forks, are swiftly resolved during the maturation of Okazaki fragments[44]. It is worth noting that contrary to observations in chicken lymphoblastoid DT40 cells and human cancer cell lines[45],

our findings demonstrate that transcription inhibition did not elicit replication timing shifts in *Ctnna2* or *Nrxn1* (Fig. S5E). This observation suggests that the influence of transcription on replication timing switching is likely locus-dependent, which further underscores the complexity of the interplay between transcription and replication dynamics in governing late-replicating TRCs.

Our investigation into the *Ptn* locus has shed light on a previously unidentified factor that governs or licenses gene loci to assume the role of RDCs. This process seems to be particularly relevant during cell fate transitions. Chromatin remodelers play a crucial role during these transitions by transforming silent or compact chromatin into a more accessible state that facilitates the activation of transcription-required trans elements[46]. We postulate that this chromatin remodeling process plays a licensing role in designating gene loci as RDCs. Notably, chromatin remodeling is closely intertwined with changes in replication timing[47,48] Consequently, we speculate that RDC loci may undergo shifts in replication timing during the transition from embryonic stem cells to neural progenitor cells (ES-to-NPC transition). Supporting this notion, the top ten robust RDCs are predominantly situated in regions exhibiting subtle changes in replication timing between ES cells and NPCs (Fig. S6). While it remains speculative, we tentatively propose that a specific chromatin state confers the license for RDC formation. Identifying the precise nature of this specific chromatin state falls beyond the purview of this article. However, such an endeavor, when pursued systematically, could potentially unveil a novel mechanism underlying TRC-mediated genome instability, offering valuable insights into the intricacies of this phenomenon.

Statistics and Reproducibility

Unless specified otherwise, statistical power in this study was determined using a nonpaired, two-tailed Student's T-test. When evaluating changes in DSB density across RDCs (Fig. 3), we analyzed 137 RDCs that exhibited more than 5 junctions. For assessing the difference

between head-on and codirectional DSB density (Fig. 4), our analysis encompassed 85 RDCs with over seven junctions. In calculating RDC DSB density for *Ctnna2* and *Nrxn1* (Fig. 5 and 6), we ensured a minimum of three replicates for each experiment. The individual experiments are represented as black dots in their respective figures. The determination of replication timing switching was performed using the R script "SwitchRT" to establish statistical significance.

ACKNOWLEDGEMENTS

This work is supported by the Helmholtz Young Investigator grant and an ERC starting grant BrainBreaks to P-C W and NIH grant R01CA270335 and GM083337 to D. M. G.. Our gratitude goes to D. M. G. for generously providing the F121-9 ES cell lines (Vozas A et al., bioRxiv), derived from paternal cell lines originating in Mitch Guttma's laboratory. V. I. acknowledges the DKFZ international Ph.D. program for a travel grant. B.D. received support from a Chinese Science Council scholarship, and J. F. was funded by a Heinrich F.C. Behr-Stiftung scholarship. L. C. was supported by an Erasmus scholarship and a scholarship from Bologna University. We extend our appreciation to Boston Bio Edit, Michael Dill, and Duncan Odom for their valuable language editing contributions. We also express our gratitude to the members of the Wei and Gilbert labs for their insightful discussions, and to the NGS Core Facility at the German Cancer Research Center (DKFZ) for their exceptional sequencing services.

AUTHOR CONTRIBUTIONS

V.I., L.Z., and P.-C. W. designed research. V.I., L.Z., A.V., M. G., J. F., G. D. M., and B. D. performed research. A.V., D. M.G. contributed new reagents. S. A. and A. A. designed the computational analytic tools. V.I., L.Z., P.-C. W. and L.-C. W. analyze the data; P.-C. W wrote the paper. V.I., L.Z., A.V., G. D. M., B. D., P.-C. W., A.V and D. M. G. helped polish the paper.

COMPETING INTERESTS

The authors declare no conflict of interest.

MATERIALS & CORRESPONDENCE

Resources, reagents, and protocol inquiries should be directed to the corresponding author (p.wei@dkfz-heidelberg.de). Oligos and commercially available reagents used in this article are detailed in Table S2. Commercially available reagents are not subjected to sharing. Mouse embryonic stem cell lines require signing a material transfer agreement (MTA) with the corresponding author's institute and other source parties. Inquiries for F121-9/CAST ES cells should be directed to David Gilbert (gilbert@sdbri.org). The raw and processed data generated for this manuscript were deposited at Gene Expression Omnibus (GSE233842, and GSE240612). Codes and bioinformatic analyses pipelines were available at BrainBreaks GitHub (https://github.com/orgs/brainbreaks/repositories).

METHODS

The convolutional neural network for termination zone prediction

While the ideal approach for this determination would involve Okazaki-sequencing (OK-seq), unfortunately, OK-seq datasets are unavailable for mouse NPCs. Thus, we employed convolutional neural network to determine the termination zones from the available 16-fraction Repli-seq datasets ([25] and GSE137764), leveraging the insights it acquired from the OK-seq datasets.

The network comprises three distinct phases. As fork direction is determined by connecting an initiation zone (IZ) to its nearest termination zone (TZ), the initial phase involves identifying genomic regions that contain defined IZ and TZ features. This step was exclusively conducted utilizing OK-seq datasets generated using mouse ES cells ([24] and GSM3290342). To predict the termination zones (TZs), we applied the OKseqHMM R package[49] to the mouse ES cell OK-seq dataset. To determined initiation zones, we fitted a LOESS model[24] to replication fork directionality (RFD – a metric used in OK-seq paper) values from mouse ES cell OK-seq data. IZs that are flanked by two TZs and with RFD value crossing zero (going from negative values to positives) were preserved as reference IZs.

The second phase involves generating training data. Initially, we select training regions that encompass at least one initiation zone (IZ), with a neighboring termination zone (TZ) situated nearby. This ensures that the region contains a complete replication fork direction. Notably, we observed instances where replication fork directions extended beyond 1.5 Mb in length in various areas. Consequently, the training area must exceed 1.5 Mb in size. To address this, we opted to enhance the dataset by introducing additional training samples produced from downscaled Repli-Seq and IZ data by a factor of two along the genomic axis, merging two adjacent 50k bins into 100k bins. Subsequently, we configured the training data to comprise 30 bins each, effectively spanning 3 Mb in each training area. This adjustment extended the coverage beyond the 1.5 Mb threshold.

Following this, the entire genome was partitioned into 1.5 Mb regions (30 bins of 50 Kb each) and larger 3 Mb regions (30 bins of 100 Kb each), and data from both segmentation strategies were unified into a cohesive dataset. This approach increased the total number of training data and allowed to cover longer replication forks. Overall, 2806 training samples were generated using a 0 x 24 stride. This approach created training data that share 6 bins with the upstream and downstream region and 24 bins that are not shared. The reason for overlap is

that when network is trained only 24 non-shared bins are predicted but the network is still able to make decisions based on data beyond that region (+/- 6 bins). With this we predict only non-overlapping regions but look at neighboring data. These samples included 763 instances with a single termination zone, 633 with two termination zones, 209 with three termination zones, 30 with four termination zones, and six with five termination zones.

The third phase is to train the model with defined training data. We choose a fully-convolutional deep learning model to handle varying genome sizes. To enhance sensitivity to various scale of replication forks we utilized single inception module[50] with dimension reduction (5x5, 7x5, 9x5 filters and max-pool layer) followed by three convolutions with 1x1, 1x1 and 1x9 filters to bring dimensions in accordance with IZ labels. Training was executed using Adam optimizer to minimize Dice loss function. A mini-batch of size 32 was used with learning rate 0.001 and dropout 90%. Activation functions within each layer was the scaled exponential linear unit (SELU) activation function, and the final output layer was a sigmoid activation function. To assess the model's performance, a 10-fold cross validation was employed. It was implemented using the Keras and Tensorflow packages within the R programming environment. Further details and the associated code are accessible on our Github https://github.com/brainbreaks/DSB_Paper.

Cell culture

If not indicated otherwise, we used genotype-matched mouse ES cell-derived NPC cell lines that were Xrcc4-/-p53-/- (Tena et al., 2020) for the described experiments. The Xrcc4/p53-deficient mouse embryonic stem cells were cultured in DMEM medium supplemented with 15% ES cell-grade fetal bovine serum, 20mM HEPES, non-essential amino acids, 100 U/ml Pen/Strep and glutamine mixture, 0.1mM beta-mercaptoethanol and 1000 U/mL ESGRO recombinant mouse leukemia inhibitory factor (LIF) on a monolayer of confluent irradiated mouse fibroblasts. For activating Ptn expression, we used Xrcc4+/+p53+/+ ES cell lines with

the F121-9/CAST genetic background. The F121-9/CAST hybrid mouse ES cells were cultured in 2i medium (50% DMEM/F12, 50% NeuralBasal, 0.5X N2 supplement, 0.5X B27 supplement, BSA, ten mM PD03259010, ten mM CHIR99021, 200 mM Glutamine, 11.9M Monothioglycerol, and 1x10^6 U/ml LIF). Xrcc4-/-p53-/- ES cell-derived NPCs were generated as described above (Tena et al., 2020; Ying et al., 2003). In brief, Xrcc4/p53-deficient mouse ES cells were plated on laminin/poly-L-ornithine-coated plates and cultured in N2B27 medium (50% DMEM/F12, 50% NeuralBasal, 1% modified N2 supplement, 2% B27 supplement without RA, 1X Glutamax) for seven days. Cells were then passaged to laminin-coated plates and culture in NBBG medium (NeralBasalA, 2% B27 supplement without RA, 0.5 mM Glutamax, 10ng/ml human EGF, 10ng/ml mouse FGFb) for another five to six days. For LAM-HTGTS experiments, ES cell-derived NPC cells were treated with or without aphidicolin with indicated concentration for 72 hours, and the aphidicolin concentration was further reduced to half for another 24 hours. For the F121-9/CAST ES cell experiments, ES cells were treated with 0.1 uM aphidicolin alone or in combination with doxycycline for 96 hours. For ATR inhibition experiments, we treated Xrcc4-/-p53-/- ES cell derived NPCs with 0.3 uM APH with or without 5 uM VE-821 for 72 hours, and the concentration of both drugs were reduced to half for the next 24 hours.

Generate a/Δ and ape/ Δ ES cell lines

For generating the a/Δ ES cell clones, we designed a single sgRNA upstream of the target gene's cis regulatory element and another downstream of the last exon. The cis regulatory elements of *Ctnna2* and *Nrxn1* genes were defined using H3K4me, H3K9ac, and H3K27ac enrichment data (ENCODE 3 ChIP-seq, UCSD/Ren). After selecting individual ES cell clones, we cultured them and conducted genotyping PCR. To generate ape/Δ ES cell clones, we designed sgRNA pairs to flank the entire cis regulatory regions of the *Ctnna2* gene and another pair for the *Nrxn1* gene. The deletion of the b allele and the *cis* regulatory elements in the *a* allele was confirmed through Sanger sequencing and Taqman qPCR.

LAM-HTGTS

Libraries were prepared following established protocols[20,21] and then sequenced on an Illumina MiSeq platform. Reads from demultiplexed FASTQ files were aligned to the mm10/GRCm38 genome assembly using Bowtie2 and further processed via the HTGTS pipeline (https://github.com/brainbreaks/HTGTS). To trigger bait on chromosomes 5, 6, 8, 12, or 17 in F121-9/CAST ES cells or Xrcc4-/-p53-/- ES cell-derived NPCs, we nucleofected 5 million cells with 5 µg of spCas9/sgRNA-expression plasmids (pX330-U6-Chimeric_BB-CBh-hSpCas9, Addgene #42230). Specific sgRNA sequences targeting bait locations (see Table S2 for reagents and resources) were cloned into separate plasmids accordingly. Utilizing appropriate reagents for ES cell-derived NPCs (VPG-1004, Lonza) or ES cells (VPH-1001, Lonza), we delivered the Cas9/sgRNA-expression plasmids using the Nucleofector 2b machine. Cells were collected 96 hours post-nucleofection for genomic DNA extraction and subsequent LAM-HTGTS analysis.

Recurrent double-stranded break clusters calling

Libraries used in this article

For RDCs detection, we combined samples from published data[20,21] (GSE106822 and GSE74356. 59 APH treated and 59 DMSO control) and newly generated samples from Chr5, 6, 8, 12, and 17 baits in Xrcc4-/-p53-/- ES cell-derived NPCs (GSE233842). APH-treated inter-chromosomal and intra-chromosomal DSB were used for the final list of RDCs. Because we observe priming strand bias at Dtel and Dcen on the bait chromosomes, analyses that involve separating LAM-HTGTS DSB directionalities are performed using inter-chromosomal DSB. The terminology “junction” and “DSB” are interchangeable in the context of LAM-HTGTS experiments.

Data clean-up

To eliminate DSBs that originated from resection around the viewpoint bait DSB, junctions from the viewpoint bait region (±6Mb, 2690102 junctions) and bait off-target regions (±50Kb, 58931 junctions) were removed. Also, duplicate reads (176774 junctions) and reads with multiple alignments (292903 junctions) were removed. Additionally, intra-chromosomal junctions from libraries where the bait chromosome was genetically modified were excluded from the subset used for RDC calling (270332 junctions). The final dataset contained 1116629 reads distributed across four conditions: APH-Inter (517947), APH-Intra (152146), DMSO-Inter (365270), and DMSO-Intra (81266).

Offtarget calling

Junctions were extended 150bp in the opposite direction from translocated prey, and only junctions that overlapped with opposite-direction junctions were kept so that there was an equal amount of centromeric and telomeric-oriented junctions. A pileup was calculated from the resulting junctions, and a Poisson distribution (mean=2) was used to calculate the significance for each interval. Continuous significant regions (p-value<0.01) were further filtered to contain more telomeric translocations upstream from the off-target site and more centromeric translocations downstream (Fisher exact test, p-value<0.01)

RDC calling

Junctions were extended 50Kb symmetrically in both directions, and pileup was calculated for telomeric-only, centromeric-only, and all junction orientations. A negative binomial model for estimating the expected pileup value for each chromosome/condition/junction-orientation triplet was derived, and a p-value was calculated for each pileup value concerning model expectation. Regions with a p-value below 0.01 joined (maximal gap 10Kb) to create seeds. These seeds were further joined with other seeds (maximal gap 100Kb) to form islands. Islands are extended up and downstream to include regions below 0.1 significance. Overlapping orientation-specific islands are further joined to form an initial RDC list that is further filtered to contain at least 100Kb below 0.01 p-value and be of at

least 300Kb in length when considering extended regions. The RDC-calling algorithm can be found here (https://github.com/brainbreaks/DSB_Paper).

LAM-HTGTS Library normalization

The LAM-HTGTS library size, DSB recovery rate, and bait induction efficiency exhibited variation across low to high aphidicolin-treated ES cell-derived NPCs. Fundamentally, aphidicolin treatment does not influence DNA break induction by CRISPR/Cas9. Elevating aphidicolin dosage augments the genome-wide DSB count, encompassing both DSBs at RDCs, with a corresponding reduction in "bait" size. To maintain the phenotype wherein increased replication stress leads to more DNA breaks, we standardized the library size to established off-target sites where DSB frequency remains unaltered by aphidicolin concentration, for each utilized bait in aphidicolin-dosage experiments. The off-target sites employed for normalization include chr2:165364620-165369885 for the chr5 bait, chr4:141559295-141560757 for the chr6 bait, chr14:22764896-22769859 for the chr8 bait, and chr12:111040435-111040503 for the chr12 bait. DSB counts from varying concentrations were adjusted based on the off-target weighting. The "off-target-normalized libraries" were utilized to determine relative DSB enrichment.

To standardize DSB counts between RDCs (Fig. 2C and 3A), the DSBs identified in DMSO, 0.2, 0.3, 0.4, and 0.6 M aphidicolin-treated cells were aggregated for each RDC (ΣRDC). ΣRDC was divided by five to calculate the mean DSB count. The value for each RDC, as illustrated in Figure 2D, was determined using the equation below. In this context, only DSBs identified through inter-chromosomal bait were considered for analysis. PRDC signifies DSB counts per treatment condition per RDC. The same equation was applied to ascertain DSBs in genes without RDCs.

Relative DSB count (Fig. 2D) = $\log2[P_{RDC}/(\Sigma RDC/5)]$

For normalizing DSB counts in the head-on and co-directional collision experiments (Fig. 3C), we only used the DSBs on the non-viewpoint chromosome. The bait preferentially recovers 15-25% more downstream DSBs at the break site chromosome than the upstream. Using viewpoint chromosome DSB resulted in overrepresention of the centromeric DSB end when the bait had a centromeric orientation. The bait that had a telomeric direction resulted in overrepresentation of the telomeric DSB end. The bias due to bait DSB end orientation on the viewpoint chromosome was as large as 20%. DSB end recovery bias did not present on the non-viewpoint chromosome.

For Figure S1D, the distance between Dcen and Dtel were roughly determined by the horizontal interval between the highest pit between Dcen and Dtel peaks.

GRO-seq

GRO-seq libraries were prepared as described[20,22]. For each GRO-seq experiment, 5 – 10 million ES cell-derived NPC nuclei were isolated for global run-on analyses. At least two technical replicates per experiment condition were performed. GRO-seq FASTQ files were aligned to the genome build mm10/GRCm38 through Bowtie2 and processed as described[20].

DRIP-seq sample preparation, sequencing, and peak calling

ES cell-derived NPCs that were treated with DMSO were used for DRIP-seq experiments. We follow the published protocol[51] with minor modifications. Two independent DRIP experiments for RNaseH-treated or untreated samples were performed. ES cell-derived NPCs at 75-80% confluency were detached by incubating with accutase at 37°C for eight minutes. Single-cell suspension was collected, and cells were pelleted down by centrifugation at 1,000 rpm for three minutes at 4°C. Cells were washed in cold DPBS, pelleted down by centrifugation again,

and resuspended in TE buffer. Ten million cells were used for the following steps. Cells were digested with proteinase K at 56°C overnight, and purified the genomic DNA with phenol/chloroform extraction as described in the protocol. Genomic DNA was digested in a restriction enzyme cocktail (HindIII, EcoRI, XbaI, BsgRI, and SspI) to derive a DNA smear between 500bp – 2kb in size. Fragmented genomic DNA was extracted following a phenol/chloroform protocol, and 10 μg genomic DNA was used for each DRIP reaction. The genomic DNA was split to two reactions. Reaction one was treated with RNAseH to remove DNA:RNA hybrid in vitro. For RNAseH-treated controls, genomic DNA was incubated with 20 units RnaseH (M0297S, NEB) at 37°C for 4-6 hours. Reaction two did not receive RNAseH treatment and thus preserved the DNA:RNA hybrid structure. DNA purified from both reactions were proceeded for DNA:RNA immunoprecipitation (DRIP). For each DRIP, 20 μg S9.6 monoclonal antibody (affinity purified from the S9.6 hybridoma culture medium) was used for DNA::RNA pulldown. Antibody-DNA::RNA complex was then pulldown with 50 μL protein G Dynabeads (10003D, Thermal Fisher Scientific). DNA::RNA complex was eluted, and the enrichment was validated by quantitative PCR at the known R-loop loci (beta-Actin, EIF5alpha) and the non-R-loop loci (chr1). DRIP-seq libraries were performed as described and were sequenced on Illumina Nextseq (75 bp single end). After sequencing, the adaptors were trimmed from the raw FASTQ reads, and the reads were aligned to mm10/GRCm38 through Bowtie2. The SAM files were transformed to BAM, sorted, and subject to MACS2 peak calling. We used the default MACS2 function, treating two DRIP samples as “treatment” and one input sample as “control.” Peaks smaller than 250 bp, with fold induction less than 10, and p or q value greater than 10^{-10} were eliminated from downstream analyses. In total, 8822 peaks were defined in the non-RnaseH-treated samples, with 8491 remaining significantly enriched in RnaseH-treated samples.

Repli-Seq sample preparation, sequencing, and analyses

Two fractions of Repli-seq were generated and analyzed as described[52]. In brief, 5 million ES cell-derived NPC cells were plated three days before BrdU incorporation. Cells were treated with 100uM BrdU for 2 hours, and cells were washed with cold DPBS 10 times to remove loosely-attached cells and dead cells. Cells were detached with Accutase, spun down, and resuspended in 2.5 ml ice-cold PBS plus 1% (vol/vol) FBS. Cells were fixed by adding 7.5 ml ice-cold 100% ethanol dropwise. The fixed cells were either stored at -20°C until sorting or directly stained with propidium iodide. Cell pellets were washed with ice-cold 1% (vol/vol) FBS in PBS once and resuspended in 0.5 ml of PBS/1% (vol/vol) FBS/propidium iodide/RNase A solution as described. The final cell suspension concentration should be 4×10^6 cells/ml. Cells were filtered through a 35 µm nylon mesh and directly to FACS. During the sorting, early or late replicating ES cell-derived NPCs were separated into two fractions; for each early and late fraction, 120,000 cells were sorted. Repli-seq libraries were generated and analyzed as described[25,52]. The cells were pelleted, and genomic DNA was extracted as described. DNA was fragmented with a Covaris S220 Focused-ultrasonicator to 200-bp average fragment size. Libraries were constructed using the NEBNext Ultra DNA Library Prep Kit before BrdU pulldown, indexing, and amplification. Before submission, the libraries were further purified with AMPure XP beads and underwent quality control. After pooling, libraries were sequenced on Illumina Nextseq (75 bp single-read) or NovaSeq (100bp single-read). After sequencing, the adaptors were trimmed from the raw FASTQ reads, and the reads were aligned to mm10/GRCm38 through Bowtie2. The significance between Repli-seq samples was determined by SwitchRT (https://github.com/ClaireMarchal/SwitchRT).

FIGURE LEGENDS

Figure 1. The convolutional neural network and the determination of replication directions. (A) The diagram presents a convolutional neural network designed for identifying replication termination zones based on OK-seq and Repli-seq data. Further information on the

selection of training regions can be found in the Methods section. On the left, four example training regions display Replication Fork Directionality (RFD) signals obtained from OK-seq and the signal from 16-fraction Repli-seq data. The training process was conducted chromosome by chromosome, employing a leave-one-out cross-validation strategy. Chromosome X was excluded from the training due to the absence of Repli-seq data. **(B)** Figure shows the precision and performance results of convolutional neural network training. A recall and precision curve and a receiver operating characteristic (ROC) curve are shown. **(C)** A histogram displays the distribution of lengths for 5844 replication forks as determined by the convolutional neural network. The average fork length is indicated by a red line.

Figure 2. RDC DSB distribution at the corresponding replication fork directions. (A) An illustration demonstrates an inward-moving fork direction, indicating a single-ended DNA break positioned on either a right-moving fork (depicted on the left) or a left-moving fork (depicted on the right). A grey arrow beneath each fork annotates replication direction. The centromere is denoted as "Cen," and the telomere as "Tel." DSB ends at the centromeric orientation, referred to as Dcen in the main text, are highlighted in blue, while DNA ends with a telomeric orientation (Dtel) are highlighted in pink. Newly synthesized DNA is represented by dark blue lines, with directionalities of DNA sequence for replicated arms indicated. **(B)** Multiomics figures provide information regarding genomic loci containing *Ctnna2* and *Prkg1*: relative oriented DSB density per 100kb, Dcen and Dtel at nucleotide resolution, replication fork directions determined by the convolutional neural network, high-resolution replication sequencing (plot generated using data from Zhao et al., 2020), and reference gene annotation. RDC loci are designated by grey boxes, with the color scheme for Dcen and Dtel explained in Figure 2A. Asterisks mark the summit positions of Dcen or Dtel peaks. In the replication direction track, fork start is depicted by colored circles, and fork end by vertical bars. Right-moving replication forks are pink, and left-moving forks are blue, with fork movement direction indicated by arrowheads. Repli-seq plot bin size is 50 kb. Reference genes spanning RDC loci are highlighted in green. **(C)** An illustration shows an unidirectional fork direction. **(D)** Multiomics figures depict the density and

location of oriented DSBs, replication direction, and gene annotations for genomic regions containing *Magi1* and *Large*, organized as described in (B). A purple rectangular shape marks the region of repetitive sequence in the *Large* locus. **(E)** An illustration presents an outward-moving fork direction. **(F)** Multiomics figures illustrate the density and location of oriented DSBs, replication direction, and gene annotations for genomic regions containing *Akap12/Zbtb2/Rmnd1* and *March1*. **(G)** An illustration provides a biphasic fork direction. **(H)** Multiomics figures showcase the density and positioning of oriented DSBs, replication direction, and gene annotations within genomic regions encompassing *Auts2* and *Sdk1*. Replication direction, as implied by Repli-seq data, is annotated with dark red and blue arrows (L: left-moving; R: right-moving).

Figure 3. Oriented DSB end position and density corresponded to DNA replication slowing. **(A)** The diagram illustrates the comparative density of DSBs in Dcen and Dtel, along with the direction of replication and the annotation of reference genes at the *Ctnna2* locus. NPCs were subjected to treatment with either 0.3 μM APH or a combination of 0.3 μM APH and 5 μM VE-821. The direction of replication forks and the annotation of reference genes are presented as explained in Figure 2B. Asterisks indicate the peak summit positions of Dcen or Dtel peaks. **(B,C)** The diagram shows the comparative density of DSBs, replication direction, and gene annotation for *Magi1* **(B)** and *Large* **(C)** loci. NPCs were subjected to the same treatment as detailed in (A). **(D)** The diagram presents the density of DSBs, nascent transcripts (GRO-seq), replication timing, and replication forks in the vicinity of the *Ctnna2* locus. The DSB density at the *Ctnna2* locus in cells exposed to different dosages of aphidicolin was shown. In the GRO-seq data, the upper panel illustrates nascent transcripts from centromeric to telomeric direction, while the lower panel depicts transcripts from telomeric to centromeric directions. *Ctnna2* encodes a transcript in the telomeric to centromeric direction. **(E)** Figure shows box plots presenting the relative DSB count of 137 RDCs. Each RDC showcases the log2-scaled off-target-normalized DSB count divided by the mean DSB count (see Methods). Significance

levels are denoted as follows: *P<0.05, ***P<0.001, ****P<0.0001, while "Ns" indicates results that were not found to be significant.

Figure 4. Head-on collision results in higher DSB density at RDCs. (A) The figure illustrates the density of DSBs within RDCs, along with nascent transcription activities (GRO-seq), replication timing, and a map of replication forks in the vicinity of the *Tenm3* locus. The structure of the figure adheres to the arrangement described in Figure 3D. **(B)** A graphical representation outlines a mathematical approach for determining the proportions of head-on and co-directional DSBs. The upper portion depicts a designated RDC marked by a black box. Within this box, DSB density values for Dcen and Dtel are positioned. Replication forks moving rightward or leftward are denoted in blue and pink, respectively. A dashed line within the RDC box indicates a termination zone. The transcription unit is depicted as a green arrow pointing leftward. In the example given, the right half of the RDC area signifies a head-on collision, while the left part represents a co-directional collision. The lower section contains two equations employed to calculate the proportions of DSBs from Dcen and Dtel in the head-on orientation (P_{HO}) or the co-directional orientation (P_{CD}). **(C)** A box plot displays the mean and 1.5 quartiles of DSB density proportions within the head-on and co-directional compartments of RDCs. A horizontal line signifies the point where Dcen and Dtel equally contribute to the P_{HO} or P_{CD} compartment (50%). Statistical significance was determined through a nonpaired, two-tailed Student's T-test, with significance levels indicated as follows: *P<0.05, ***P<0.001, ****P<0.0001. Results not found to be significant are denoted as "Ns" (not significant).

Figure 5. Transcription inactivation abolishes RDC formation in the *Ctnna2* locus. (**A**) Visual representation of the experimental strategy for silencing transcription in *Ctnna2*. At the top, the parental NPC (a/b) actively transcribes both alleles. In the middle, a section spanning 2.2 kb upstream and 2.6 kb downstream of *Ctnna2* on the *b* allele has been removed, resulting in an Ctnna2-a/Δ ES cell-derived NPC clone. At the bottom, two Ctnna2-ape/Δ NPC clones were obtained by deleting a region encompassing 5 kb upstream and 17 kb downstream of

the transcription start site at *Ctnna2*. **(B)** Figure shows the nascent transcripts (GRO-seq, top) and the DSB distribution (LAM-HTGTS, bottom) in the vicinity of the *Ctnna2* locus. The upper GRO-seq panels display transcription in the direction from telomere to centromere (gray) or centromere to telomere (black). The region encompassing the *Ctnna2* gene is highlighted in yellow. The Y-axis represents reads per kilobase per million. The bottom panel illustrates the distribution of DSBs in and around the *Ctnna2* locus. LAM-HTGTS DSB data were normalized to 50,000 per library for plotting. Each black line indicates a DSB. +/- APH indicates treatment with or without aphidicolin. **(C, D)** DSB density assessment at the *Ctnna2* (C) and *Foxp1* (D) Gene loci in different NPCs. Intrachromosomal DSBs derived from Chr6-bait DSB libraries were used to calculate DSB density, presented as thousand DSBs per megabase. DSB density from NPCs treated with (orange) or without (blue) APH is displayed. Significance levels: *p<0.05, **p<0.005, ***p<0.0005, and ****p<0.0005, determined using a one-tailed type two T-test. N.s. indicates results that are not significant.

Figure 6. *Nrxn1* RDC requires transcription of long *Nrxn1* isoform. (A) Experimental approach to suppress transcription of the long *Nrxn1* isoform. The figure layout follows the structure outlined in Figure 4A. In Nrxn1-a/Δ ES cell-derived NPCs, a segment spanning 4 kb upstream and 1.2 kb downstream of Nrxn1 on the b allele was removed. Similarly, in Nrxn1-ape/Δ ES cell-derived NPC clones, a region comprising 1.5 kb upstream and 9.6 kb downstream of the transcription start site at Nrxn1 was deleted. **(B)** Transcript analysis and DSB position surrounding *Nrxn1* gene locus. The *Nrxn1* gene region is highlighted in yellow. The top panel indicates different *Nrxn1* gene isoforms, with a brown dashed line marking the transcription start site for the short Nrxn1 isoform. The middle panel showcases nascent transcripts (GRO-seq), while the lower panels display the position of DSB (LAM-HTGTS DSB) at and near the *Nrxn1* gene locus. DSBs were standardized to 30,000 per library for plotting. **(C, D)** Assessment of DSB density at the *Nrxn1* **(C)** and *Ptprm* **(D)** gene loci in NPCs. The format of the figures adheres to that described in Figures 4C, D. Significance levels: *p<0.05,

**p<0.005, ***p<0.0005, and ****p<0.0005, determined through a one-tailed type two T-test. N.s. denotes outcomes without statistical significance.

Figure 7. Transcription activation failed to induce RDC at the *Ptn* locus. (A) Diagram depicting the F121 allele of *Ptn* gene in F121-9/CAST ES cells. The configuration displays the endogenous *Ptn* promoter (grey), exogenous mouse PGK promoter (red), and human PGK promoter (light blue) on the F121 alleles. The six exons of *Ptn* are represented by thick dark blue lines, while transcription start sites are marked by black arrows. Green curvy lines symbolize transcripts. Alleles lacking transcription activity are indicated with a red-circle-cross. **(B)** GRO-seq data of nascent transcription activity surrounding *Ptn* locus. Panels illustrate the nascent transcription activity (GRO-seq) around the *Ptn* gene in F121-9/CAST ES cells with different promoters. **(C)** Detection of DSBs at the *Ptn* locus in ES Cells with aphidicolin treatment. The upper section presents a schematic of the LAM-HTGTS bait and its relevant position on mouse chromosome 6, with the green box indicating the LAM-HTGTS bait. Grey arrowed lines represent possible viewpoints for "prey" joining. Grey boxes depict DSBs at the *Ptn* gene. The lower part shows DSBs detected by LAM-HTGTS. **(D)** The diagram showcases the *Ptn* gene allele, where the native Ptn promoter is substituted with a TRE sequence on the F121 allele. The TRE-Ptn cell line constitutively expresses rtTA, and the presence of doxycycline initiates *Ptn* expression. **(E)** LAM-HTGTS detection of DSBs in the Iinducible *Ptn* expression context. The upper part depicts the LAM-HTGTS viewpoint and the "prey" region within *Ptn*. The lower part shows DSBs detected by LAM-HTGTS in the *Ptn* locus and the nearby area. **(F)** LAM-HTGTS detection of DSBs in comparison to *Ctnna2*. The upper segment illustrates the LAM-HTGTS viewpoint and the "prey" area in *Ctnna2*. The lower part displays DSBs detected through LAM-HTGTS in the *Ctnna2* locus and the adjacent region. For **(E)** and **(F)**, the figures present genome-wide DSBs: 15,034 (without dox/APH), 19,787 (without dox/+APH), 19,187 (0.125 μg/mL dox/+APH), 20,115 (0.25 μg/mL dox/+APH), and 18,584 (0.5 μg/mL dox/+APH).

SUPPLEMENTARY FIGURE LEGENDS

Figure S1. Using LAM-HTGTS to determine DSB end accessibility. Related to main text and Figure 2. (A) The LAM-HTGTS workflow and differentiation of centromeric and telomeric DSB ends. Within living cells, a bait DSB is introduced through the CRISPR/Cas system. The position of the DSB is indicated by a black vertical arrow located at the sgRNA targeting sequences. Upon CRISPR/Cas9 cutting, it generates two DSB ends to be used as viewpoints. These ends are highlighted as green rectangles. The terms "Cen" and "Tel" correspond to centromere and telomere, respectively. These centromeric and telomeric DSB ends interact at the viewpoint and connect with "prey" DSBs located on the same or different chromosomes. This process leads to an equivalent frequency of translocations from the bait DSB to either the centromeric or telomeric prey DSB ends when both are available. The translocation involving prey DSBs pointing in the telomeric direction is denoted as "Dcen " or "Dtel," based on the orientation. Genomic DNA is subsequently extracted for further procedures. In a test tube, the genomic DNA undergoes fragmentation via sonication. A single primer (indicated by a red arrowhead) positioned upstream of the viewpoint DSB is employed for linear-amplification mediated PCR (LAM-PCR). This technique maintains the strand orientation of prey sequences and DSB orientation. The subsequent steps of LAM-HTGTS library preparation adhere to the protocol established by Hu et al[27]. In our approach, we differentiate DSBs with a "Dcen " tag from those with a "Dtel" tag, as these represent DSBs with open ends associated with centromeric sequences or telomeric sequences, respectively. **(B-D)** Cartoons illustrate four possible outcomes depicting the position and density of Dcen and Dtel as unveiled by LAM-HTGTS. Top panel shows four sample chromosomes, each featuring an attachment of a Dcen and a Dtel end to both the centromere and telomere. The lower panel presents (1) the positioning of individual Dcen and Dtel DSBs, considering a total of 50 ends from each orientation, (2) DSB density per 100kb for Dcen ends, (3) DSB density per 100kb for Dtel ends, and (4) a view indicating the overlap between Dcen and Dtel peaks. In genomic regions where the distance between Dcen and Dtel ends is under 100 kb, the Dcen and Dtel peaks exhibit a

clear overlap **(B).** When Dcen and Dtel are separated by more than 100 kb in a genomic region, a gap appears between the two peaks, and they do not overlap **(C).** In a genomic region where only either Dcen or Dtel ends are present, the overlapping view would display only a single DSB density peak **(D).**

Figure S2. The interval distance between Dcen and Dtel peaks in inward-moving fork directions. Related to Figure 2. (A) A scatter plot depicting the interval distance between Dcen and Dtel peaks. The graph includes 59 RDCs featuring twin-peak signatures that move inward. Each dot on the plot represents a gene-containing RDC. The annotation highlights the two RDCs with the most significant Dcen-Dtel interval. The two RDC-containing genes with the longest interval are labeled. A red line signifies the median length of the Dcen and Dtel interval.

Figure S3. Validation of clone Identity for Ctnna2-a/Δ and Ctnna2-ape/Δ Cells. Related to Figure 5. (A) Clone identity verification for Ctnna2-a/Δ and Ctnna2-ape/Δ ES cells. At the top, the figure illustrates the design for CRISPR/Cas9-mediated deletion at the *Ctnna2* locus. sgRNA_Ctnna2_ds1 and sgRNA_Ctnna2_up1 target downstream (blue) and upstream sequences (yellow) of the *Ctnna2* gene, respectively. Two purple arrows represent PCR primers that amplify DNA segments upstream and downstream of sgRNA_Ctnna2_ds1 or sgRNA_Ctnna2_up1. The expected size of the PCR product resulting from the deletion is 443 bp. At the bottom, Sanger sequencing results of the PCR product from the *b* allele amplified using Ctnna2-a/Δ genomic DNA are presented. A black dashed line marks the deletion junction. The original unmodified sequences at the junction are shown in yellow and blue. The sequences immediately upstream and downstream that were deleted in this clone are highlighted in black. **(B)** Sanger sequencing results of deletional junctions for the *a* allele of Ctnna2-ape/Δ clones. The expected size of the PCR product resulting from the deletion is 510 base pairs. The original unmodified sequences belonging to the upstream (yellow) and downstream (blue) of the sgRNA pairs are depicted. (C) TaqMan qPCR analysis of *Ctnna2* allelic number. The top section illustrates a simplified *Ctnna2* locus, where the number of exons

and distances between them are not proportionate to the actual locus. The positions of TaqMan qPCR probes targeting the middle of *Ctnna2* (allelic probe) and the promoter-enhanced region (Prom.enh probe) are indicated. The bottom part presents a bar chart displaying the results of TaqMan qPCR analysis, showing the allelic ratio for both the *Ctnna2* gene body (blue) and the *Ctnna2* promoter/enhancer region (orange). The allelic ratio is normalized to the *Apob* gene (chr12:8010487-8010560). The corresponding *Ctnna2* clones used in this analysis are listed below the bar chart. Mean and standard deviations are represented in the chart. "N.d." indicates results that were not detected.

Figure S4. Validation of clone identity for Nrxn1-a/Δ and Nrxn1-ape/Δ ES cells and replication timing profiles at *Ctnna2* and *Nrxn1* loci. Related to Figure 6. (A) Validation of clone Identity for Nrxn1-a/Δ and Nrxn1-ape/Δ ES cells. The figure outlines the CRISPR/Cas9-mediated deletion approach at the *Nrxn1* locus, and it presents the Sanger sequencing results of the deletion on the Nrxn1 b allele. **(B)** Sanger sequencing results of the promoter/enhancer deletional Junction for the *a* allele of the *Nrxn1* Gene. **(C)** TaqMan qPCR analysis of *Nrxn1* allelic number. A bar chart illustrates the results of TaqMan qPCR analysis, presenting the allelic ratio for both the *Nrxn1* gene body and the promoter/enhancer area. "N.d." signifies results not detected. **(D, E)** Figures showcase the results of two-fraction replication timing sequencing around the *Ctnna2* **(D)** or *Nrxn1* **(E)** loci. The timing differences between the early and late fractions are presented on a logarithmic scale. Two technical repeats are shown for each condition. Statistical significance between clones is indicated, with *P<0.05 denoting significance.

Figure S5. *Ptn* RDC in ES Cell-Derived NPCs and *Ptn* gene expression levels in ES Cells. Related to Figure 7. (A) *Ptn* RDC in Xrcc4/p53-deficient ES cell-derived NPCs. The figure illustrates the distribution of DSBs around the *Ptn* locus in NPC. The DSBs have been standardized to 170,000 per library for figure plotting. Each black line corresponds to a single DSB. The addition of aphidicolin (APH) is indicated. **(B)** *Ptn* gene expression levels in F121-

9/CAST ES cells. Bar charts present the levels of *Ptn* transcripts from three separate experiments. RT-qPCR was conducted using primer pairs targeting exon 1 and 2, exon 3 and 4, and exon 4 and 5 of *Ptn* transcripts. The *Ptn* expression is normalized to *Gapdh*. Mean from each RT-qPCR run are provided.

Figure S6. Replication Timing of the Strongest RDC Loci in ES Cells and NPCs. Related to Discussion. High-resolution Repli-seq data obtained from mouse ES cells and NPCs[25] are depicted for the ten most prominent RDC loci (Wei 2018). The Repli-seq panels maintain the organization seen in preceding figures. Genes that encompass RDCs are delineated within yellow rectangular boxes.

Supplementary Table 1

The Excel spreadsheet encompasses information about the 152 RDCs analyzed in this study, including their respective chromosome (RDCChrom), start (RDCstart), and end (RDCend) coordinates, the associated gene names (> 100kb) that colocalize with each RDC, the corresponding RDC name (Gene (RDC) name), the coding strand of the gene (Gene Coding strand), the start (Gene start) and end (Gene end) coordinates of the indicated genes, the negative logarithm of the P value of the RDC (-log10P), the RDC group classification (Group (Wei et al., 2018 PNAS)), biphasic status of RDC (biphasic, yes or no), the number of DRIP-seq peaks within the RDC (DRIP peaks), the fork direction determined through the convolutional neural network (fork direction), and the shape of the Dcen and Dtel peaks (Peak shape).

Supplementary Table 2

The Excel spreadsheet comprises oligo sequences and reagent details employed in this research. The specific function of each oligo and reagent is indicated alongside.

REFERENCES

1. Berti, M., Cortez, D. & Lopes, M. The plasticity of DNA replication forks in response to clinically relevant genotoxic stress. *Nat Rev Mol Cell Bio* **21**, 633–651 (2020).

2. Fu, H. *et al.* The DNA repair endonuclease Mus81 facilitates fast DNA replication in the absence of exogenous damage. *Nat. Commun.* **6**, 6746 (2015).

3. McPherson, J. P. *et al.* Involvement of Mammalian Mus81 in Genome Integrity and Tumor Suppression. *Science* **304**, 1822–1826 (2004).

4. Tubbs, A. & Nussenzweig, A. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell* **168**, 644–656 (2017).

5. Ciccia, A., McDonald, N. & West, S. C. Structural and Functional Relationships of the XPF/MUS81 Family of Proteins. *Biochemistry-us* **77**, 259–287 (2008).

6. Alexander, J. L. & Orr-Weaver, T. L. Replication fork instability and the consequences of fork collisions from rereplication. *Genes Dev.* **30**, 2241–2252 (2016).

7. Dehé, P.-M. & Gaillard, P.-H. L. Control of structure-specific endonucleases to maintain genome stability. *Nat Rev Mol Cell Bio* **18**, 315–330 (2017).

8. Mukherjee, S., Wright, W. D., Ehmsen, K. T. & Heyer, W.-D. The Mus81-Mms4 structure-selective endonuclease requires nicked DNA junctions to undergo conformational changes and bend its DNA substrates for cleavage. *Nucleic Acids Res.* **42**, 6511–6522 (2014).

9. Doe, C. L., Ahn, J. S., Dixon, J. & Whitby, M. C. Mus81-Eme1 and Rqh1 Involvement in Processing Stalled and Collapsed Replication Forks*. *J. Biol. Chem.* **277**, 32753–32759 (2002).

10. Helmrich, A., Ballarino, M. & Tora, L. Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Mol Cell* **44**, 966–77 (2011).

11. Wilson, T. E. *et al.* Large transcription units unify copy number variants and common fragile sites arising under replication stress. *Genome Res* **25**, 189–200 (2014).

12. Glover, T. W., Berger, C., Coyle, J. & Echo, B. DNA polymerase α inhibition by aphidicolin induces gaps and breaks at common fragile sites in human chromosomes. *Hum Genet* **67**, 136–142 (1984).

13. Ying, S. *et al.* MUS81 promotes common fragile site expression. *Nat Cell Biol* **15**, 1001–1007 (2013).

14. Naim, V., Wilhelm, T., Debatisse, M. & Rosselli, F. ERCC1 and MUS81–EME1 promote sister chromatid separation by processing late replication intermediates at common fragile sites during mitosis. *Nat Cell Biol* **15**, 1008–1015 (2013).

15. Glover, T. W., Wilson, T. E. & Arlt, M. F. Fragile sites in cancer: more than meets the eye. *Nat Rev Cancer* **17**, 489–501 (2017).

16. Hastings, P. J., Ira, G. & Lupski, J. R. A Microhomology-Mediated Break-Induced Replication Model for the Origin of Human Copy Number Variation. *PLoS Genet.* **5**, e1000327 (2009).

17. Petermann, E., Lan, L. & Zou, L. Sources, resolution and physiological relevance of R-loops and RNA–DNA hybrids. *Nat. Rev. Mol. Cell Biol.* **23**, 521–540 (2022).

18. Hamperl, S., Bocek, M. J., Saldivar, J. C., Swigut, T. & Cimprich, K. A. Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses. *Cell* **170**, 774-786.e19 (2017).

19. Sanz, L. A. *et al.* Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Mol. Cell* **63**, 167–178 (2016).

20. Wei, P.-C. *et al.* Long Neural Genes Harbor Recurrent DNA Break Clusters in Neural Stem/Progenitor Cells. *Cell* **164**, 644–55 (2016).

21. Wei, P.-C. *et al.* Three classes of recurrent DNA break clusters in brain progenitors identified by 3D proximity-based break joining assay. *Proc National Acad Sci* **115**, 1919–1924 (2018).

22. Tena, A. *et al.* Induction of recurrent break cluster genes in neural progenitor cells differentiated from embryonic stem cells in culture. *Proc National Acad Sci* **117**, 10541–10546 (2020).

23. Wang, M. *et al.* Increased Neural Progenitor Proliferation in a hiPSC Model of Autism Induces Replication Stress-Associated Genome Instability. *Cell Stem Cell* **26**, 221-233.e6 (2020).

24. Petryk, N. *et al.* MCM2 promotes symmetric inheritance of modified histones during DNA replication. *Science* **361**, 1389–1392 (2018).

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

26. Petryk, N. *et al.* Replication landscape of the human genome. *Nat Commun* **7**, 10208 (2016).

27. Hu, J. *et al.* Detecting DNA double-stranded breaks in mammalian genomes by linear amplification–mediated high-throughput genome-wide translocation sequencing. *Nat Protoc* **11**, 853–871 (2016).

28. Frock, R. L. *et al.* Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat Biotechnol* **33**, 179–86 (2014).

29. Barlow, J. H. *et al.* Identification of early replicating fragile sites that contribute to genome instability. *Cell* **152**, 620–32 (2013).

30. Toledo, L. I. *et al.* ATR Prohibits Replication Catastrophe by Preventing Global Exhaustion of RPA. *Cell* **156**, 374 (2014).

31. Moiseeva, T. N. *et al.* An ATR and CHK1 kinase signaling mechanism that limits origin firing during unperturbed DNA replication. *Proc. Natl. Acad. Sci.* **116**, 13374–13383 (2019).

32. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322**, 1845–1848 (2008).

33. Thongthip, S., Carlson, A., Crossley, M. P. & Schwer, B. Relationships between genome-wide R-loop distribution and classes of recurrent DNA breaks in neural stem/progenitor cells. *Sci Rep-uk* **12**, 13373 (2022).

34. Park, S. H. *et al.* Locus-specific transcription silencing at the FHIT gene suppresses replication stress-induced copy number variant formation and associated replication delay. *Nucleic Acids Res* **49**, 7507–7524 (2021).

35. Julienne, H., Audit, B. & Arneodo, A. Embryonic Stem Cell Specific “Master” Replication Origins at the Heart of the Loss of Pluripotency. *PLoS Comput. Biol.* **11**, e1003969 (2015).

36. Paganelli, V., Giordano, M., Meazza, C., Schena, L. & Bozzola, M. An intragenic deletion within CTNNA2 intron 7 in a boy with short stature and speech delay: A case report. *Sage Open Medical Case Reports* **5**, 2050313X17693967 (2017).

37. Duda, H. *et al.* A Mechanism for Controlled Breakage of Under-replicated Chromosomes during Mitosis. *Dev Cell* **39**, 740–755 (2016).

38. Bhowmick, R., Minocherhomji, S. & Hickson, I. D. RAD52 Facilitates Mitotic DNA Synthesis Following Replication Stress. *Mol Cell* **64**, 1117–1126 (2016).

39. Bétous, R. *et al.* DNA replication stress triggers rapid DNA replication fork breakage by Artemis and XPF. *Plos Genet* **14**, e1007541 (2018).

40. Anand, R., Beach, A., Li, K. & Haber, J. Rad51-mediated double-strand break repair and mismatch correction of divergent substrates. *Nature* **544**, 377–380 (2017).

41. Ira, G. & Haber, J. E. Characterization of RAD51-Independent Break-Induced Replication That Acts Preferentially with Short Homologous Sequences. *Mol. Cell. Biol.* **22**, 6384–6392 (2002).

42. Mehta, A., Beach, A. & Haber, J. E. Homology Requirements and Competition between Gene Conversion and Break-Induced Replication during Double-Strand Break Repair. *Mol. Cell* **65**, 515-526.e3 (2017).

43. Pannunzio, N. R., Li, S., Watanabe, G. & Lieber, M. R. Non-homologous end joining often uses microhomology: Implications for alternative end joining. *DNA Repair* **17**, 74–80 (2014).

44. Gloor, J. W., Balakrishnan, L., Campbell, J. L. & Bambara, R. A. Biochemical analyses indicate that binding and cleavage specificities define the ordered processing of human Okazaki fragments by Dna2 and FEN1. *Nucleic Acids Res.* **40**, 6774–6786 (2012).

45. Blin, M. *et al.* Transcription-dependent regulation of replication dynamics modulates genome stability. *Nat Struct Mol Biol* **26**, 58–66 (2019).

46. Clapier, C. R. & Cairns, B. R. The Biology of Chromatin Remodeling Complexes. *Annu. Rev. Biochem.* **78**, 273–304 (2009).

47. Schwaiger, M., Kohler, H., Oakeley, E. J., Stadler, M. B. & Schübeler, D. Heterochromatin protein 1 (HP1) modulates replication timing of the Drosophila genome. *Genome Res.* **20**, 771–780 (2010).

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

49. Liu, Y., Wu, X., d'Aubenton-Carafa, Y., Thermes, C. & Chen, C.-L. OKseqHMM: a genome-wide replication fork directionality analysis toolkit. *Nucleic Acids Res.* **51**, e22–e22 (2023).

50. Szegedy, C. *et al.* Going Deeper with Convolutions. *arXiv* (2014) doi:10.48550/arxiv.1409.4842.

51. Sanz, L. A. & Chédin, F. High-resolution, strand-specific R-loop mapping via S9.6-based DNA-RNA immunoprecipitation and high-throughput sequencing. *Nat Protoc* **14**, 1734–1755 (2019).

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

Figure 1

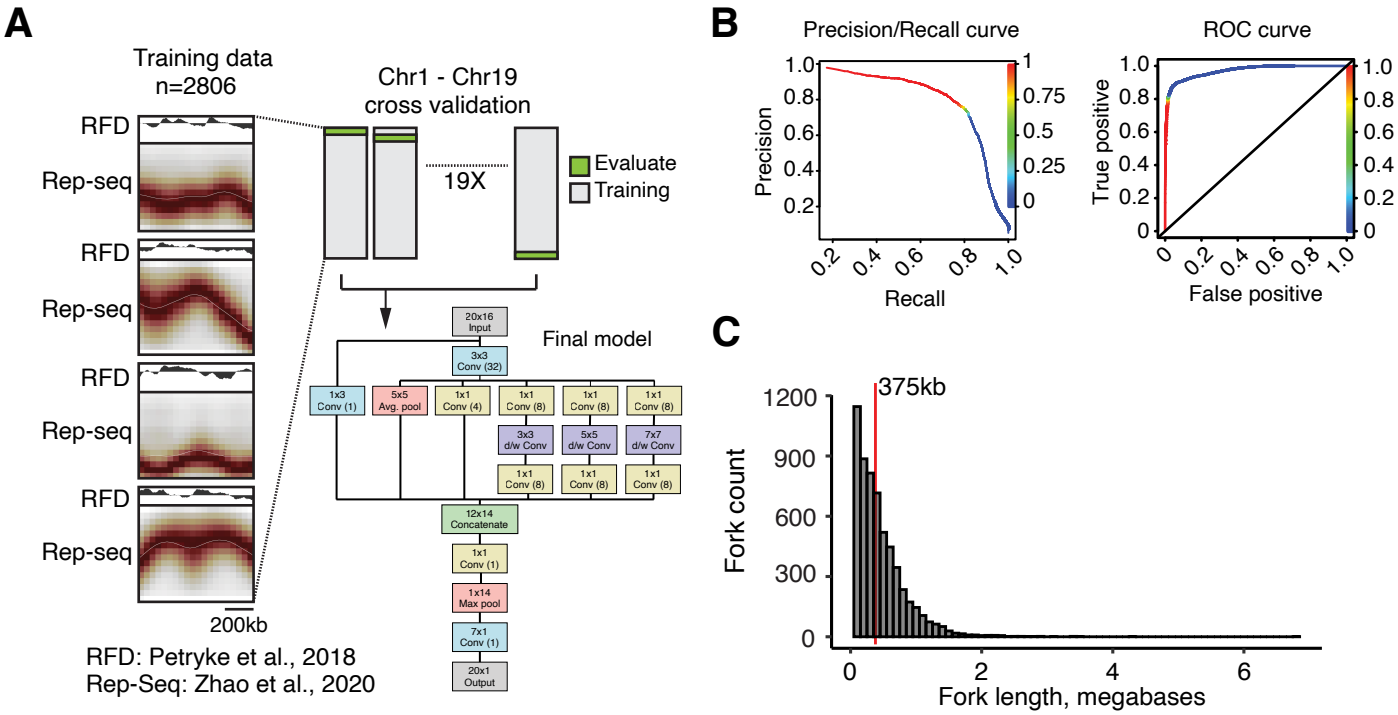

A
Training data
n=2806
RFD
Rep-seq
RFD
Rep-seq
RFD
Rep-seq
RFD
Rep-seq
200kb
RFD: Petryke et al., 2018
Rep-Seq: Zhao et al., 2020
Chr1 - Chr19
cross validation
19X
Evaluate
Training
Final model
20x16 Input
3x3 Conv (32)
1x3 Conv (1)
5x5 Avg. pool
1x1 Conv (4)
1x1 Conv (8)
1x1 Conv (8)
1x1 Conv (8)
3x3 d/w Conv
5x5 d/w Conv
7x7 d/w Conv
1x1 Conv (8)
1x1 Conv (8)
1x1 Conv (8)
12x14 Concatenate
1x1 Conv (1)
1x14 Max pool
7x1 Conv (1)
20x1 Output
B
Precision/Recall curve
Precision
Recall
ROC curve
True positive
False positive
C
375kb
Fork count
Fork length, megabases

Figure 2

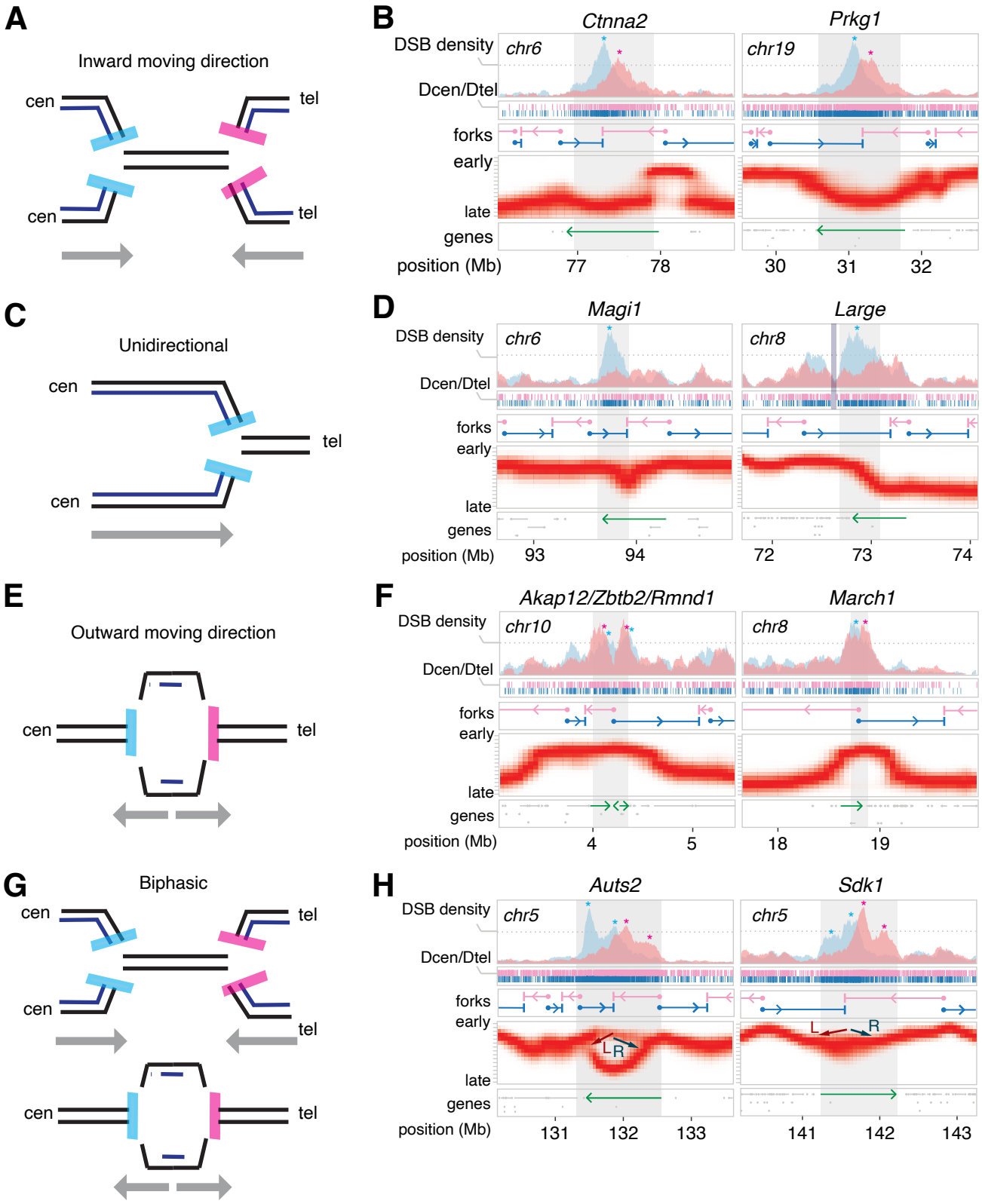

A
Inward moving direction
cen
tel
cen
tel
B
Ctnna2
Prkg1
DSB density
chr6
chr19
Dcen/Dtel
forks
early
late
genes
position (Mb)
77
78
30
31
32
C
Unidirectional
cen
tel
cen
D
Magi1
Large
chr6
chr8
93
94
72
73
74
E
Outward moving direction
cen
tel
F
Akap12/Zbtb2/Rmnd1
March1
chr10
chr8
4
5
18
19
G
Biphasic
cen
tel
cen
tel
cen
tel
H
Auts2
Sdk1
chr5
chr5
L
R
131
132
133
141
142
143

Figure 3

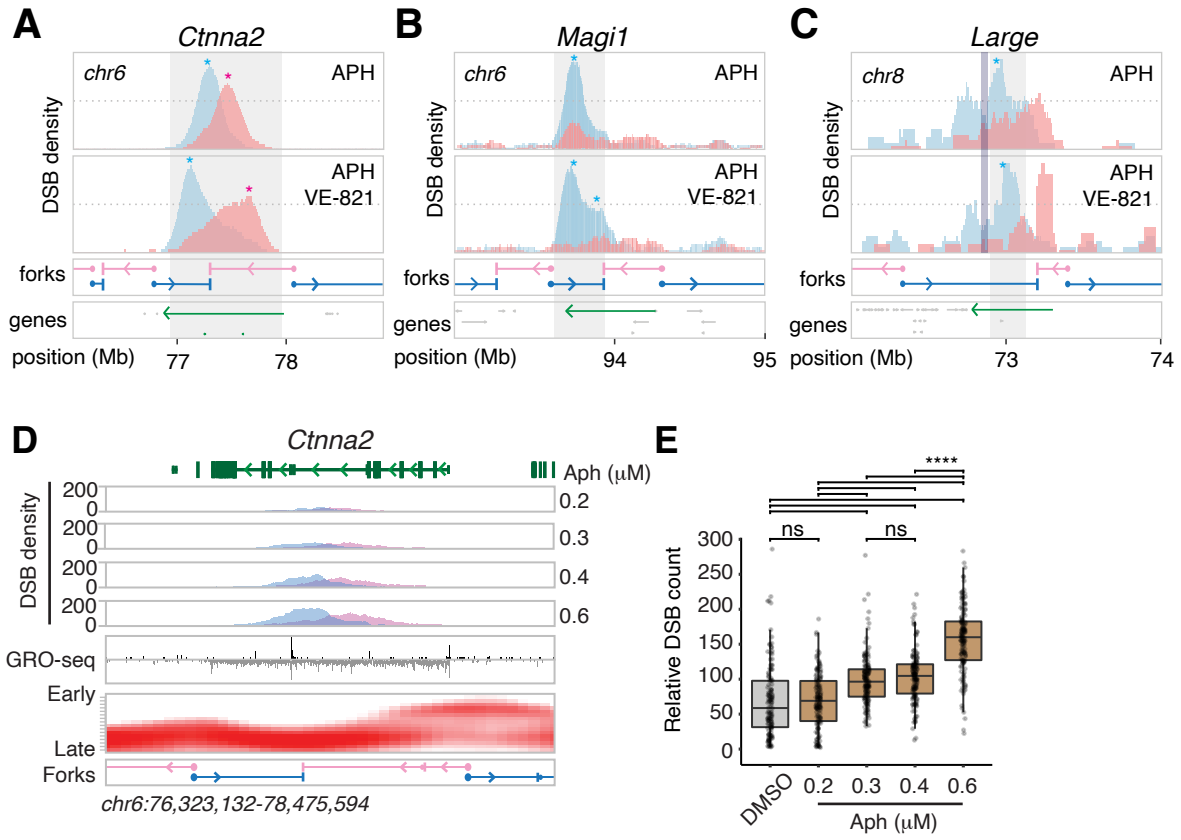

A
Ctnna2
chr6
APH
APH
VE-821
DSB density
forks
genes
position (Mb)
77
78
B
Magi1
chr6
APH
APH
VE-821
DSB density
forks
genes
position (Mb)
94
95
C
Large
chr8
APH
APH
VE-821
DSB density
forks
genes
position (Mb)
73
74
D
Ctnna2
Aph (μM)
DSB density
200
0
0.2
0.3
0.4
0.6
GRO-seq
Early
Late
Forks
chr6:76,323,132-78,475,594
E
****
ns
ns
Relative DSB count
300
250
200
150
100
50
0
DMSO
0.2
0.3
0.4
0.6
Aph (μM)

Figure 4

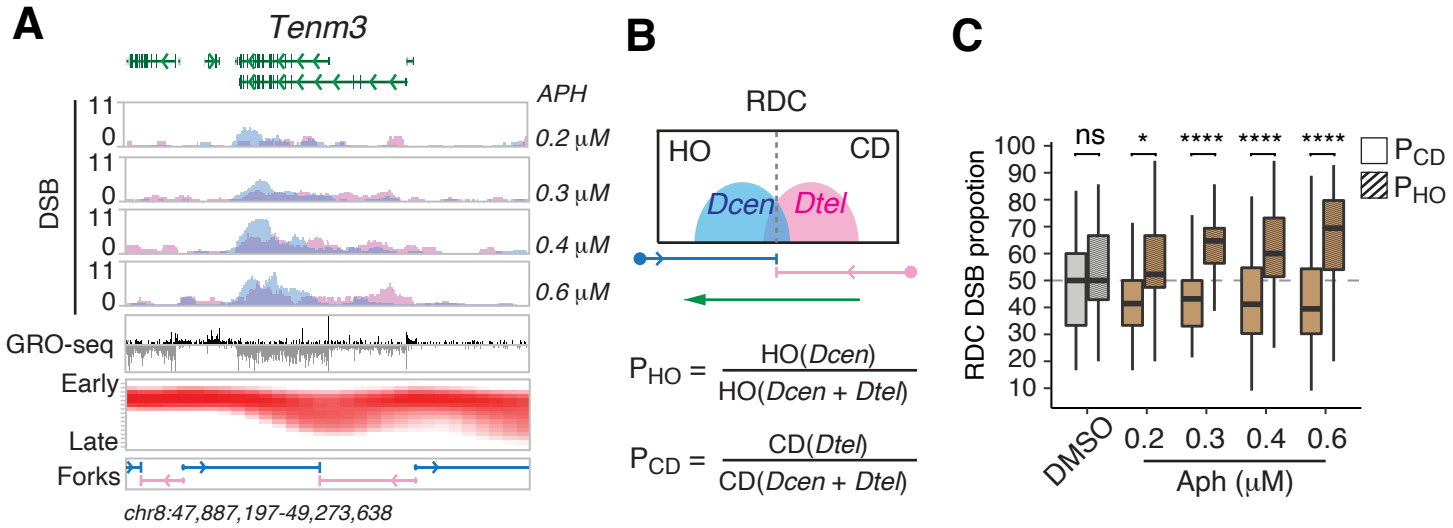

A
Tenm3
APH
DSB
11
0
0.2 μM
0.3 μM
0.4 μM
0.6 μM
GRO-seq
Early
Late
Forks
chr8:47,887,197-49,273,638
B
RDC
HO
CD
Dcen
Dtel
$P_{HO} = \frac{HO(Dcen)}{HO(Dcen + Dtel)}$
$P_{CD} = \frac{CD(Dtel)}{CD(Dcen + Dtel)}$
C
RDC DSB propotion
ns
*
****
P_{CD}
P_{HO}
DMSO
0.2
0.3
0.4
0.6
Aph (μM)

Figure 5

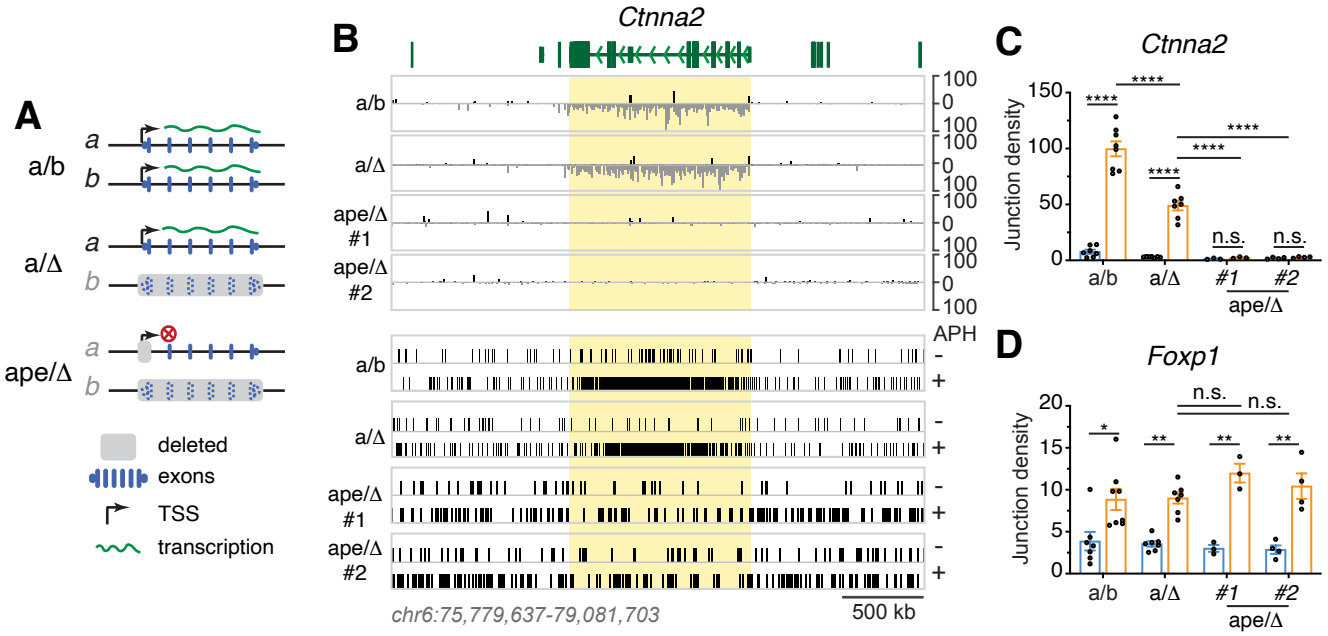

A
a/b
a
b
a/Δ
a
b
ape/Δ
a
b
deleted
exons
TSS
transcription
B
Ctnna2
a/b
a/Δ
ape/Δ #1
ape/Δ #2
100
0
100
APH
-
+
chr6:75,779,637-79,081,703
500 kb
C
Ctnna2
Junction density
150
100
50
0
****
n.s.
a/b
a/Δ
#1
#2
ape/Δ
D
Foxp1
Junction density
20
15
10
5
0
*
**
n.s.
a/b
a/Δ
#1
#2
ape/Δ

Figure 6

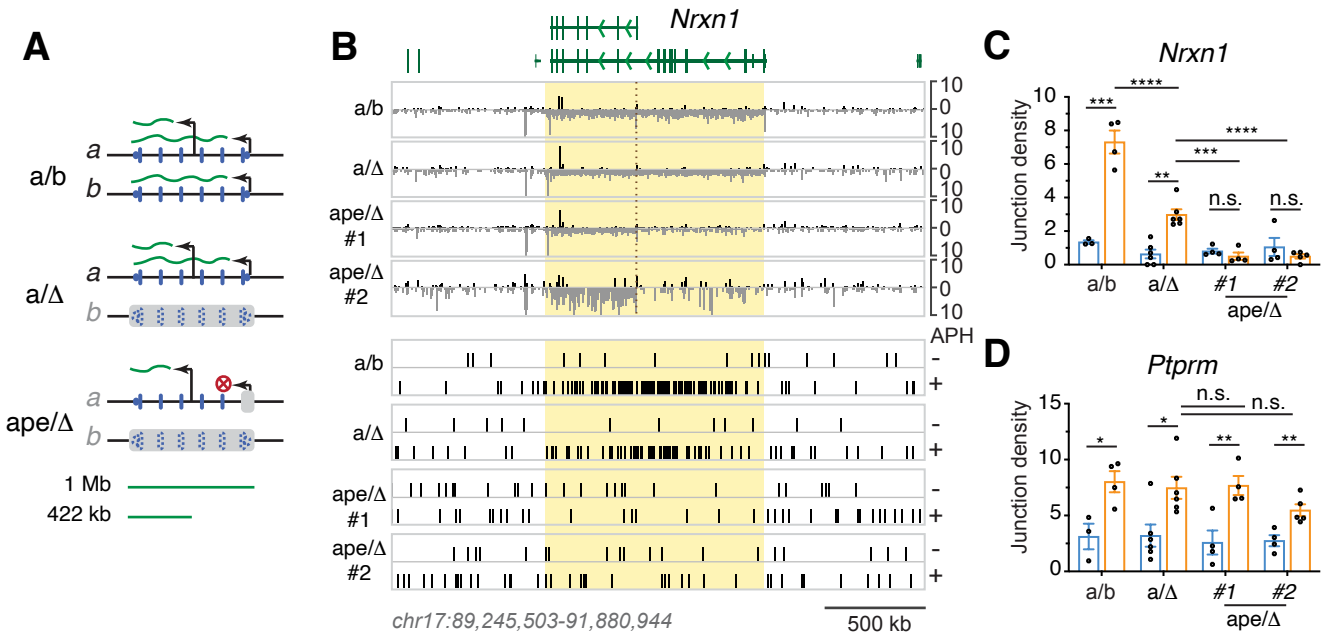
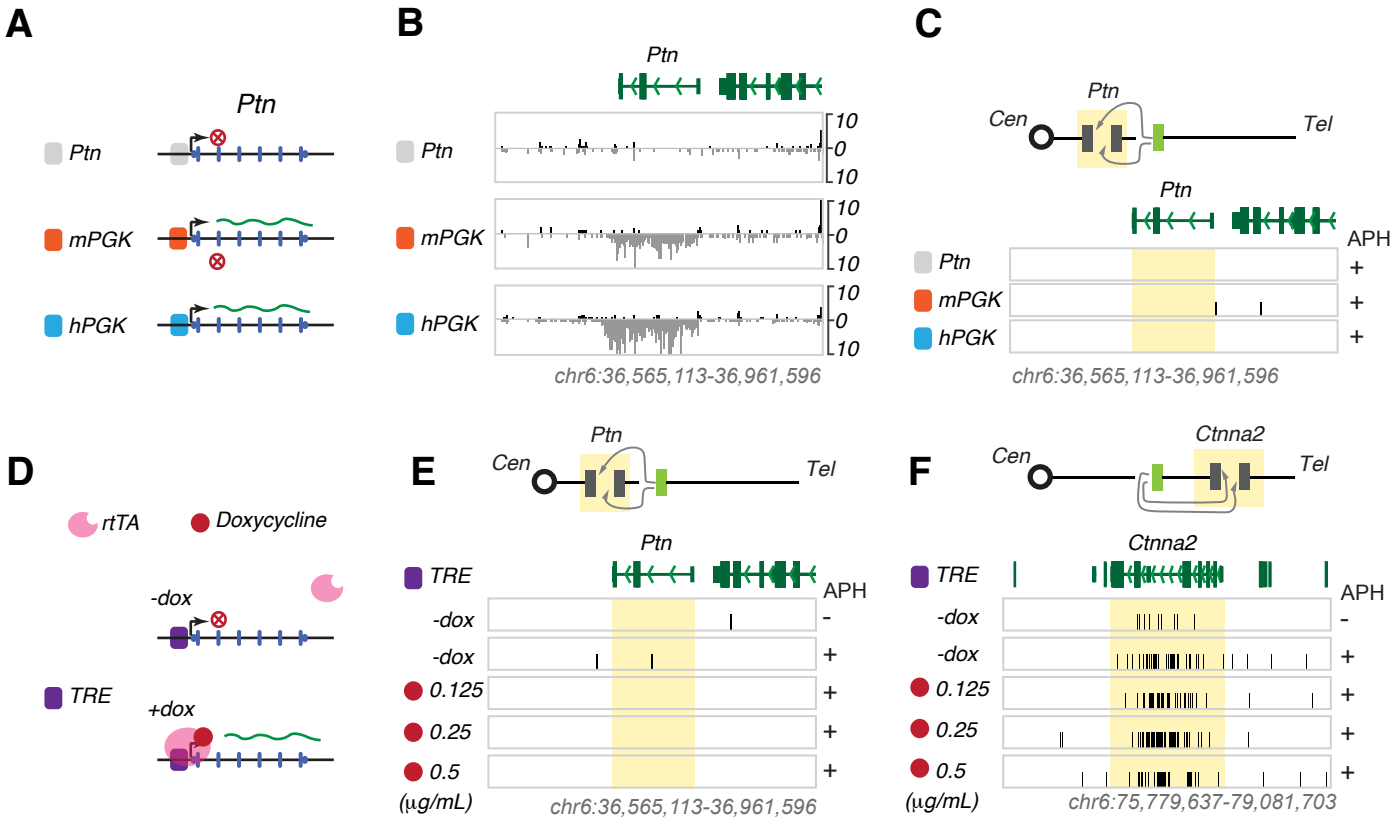

A
a/b
a
b
a/Δ
a
b
ape/Δ
a
b
1 Mb
422 kb
B
Nrxn1
a/b
a/Δ
ape/Δ #1
ape/Δ #2
10
0
10
APH
-
+
chr17:89,245,503-91,880,944
500 kb
C
Nrxn1
Junction density
a/b
a/Δ
#1
#2
ape/Δ
n.s.
D
Ptprm
Junction density
n.s.

Figure 7



Ionasz and Corazzi et al.

Figure S1

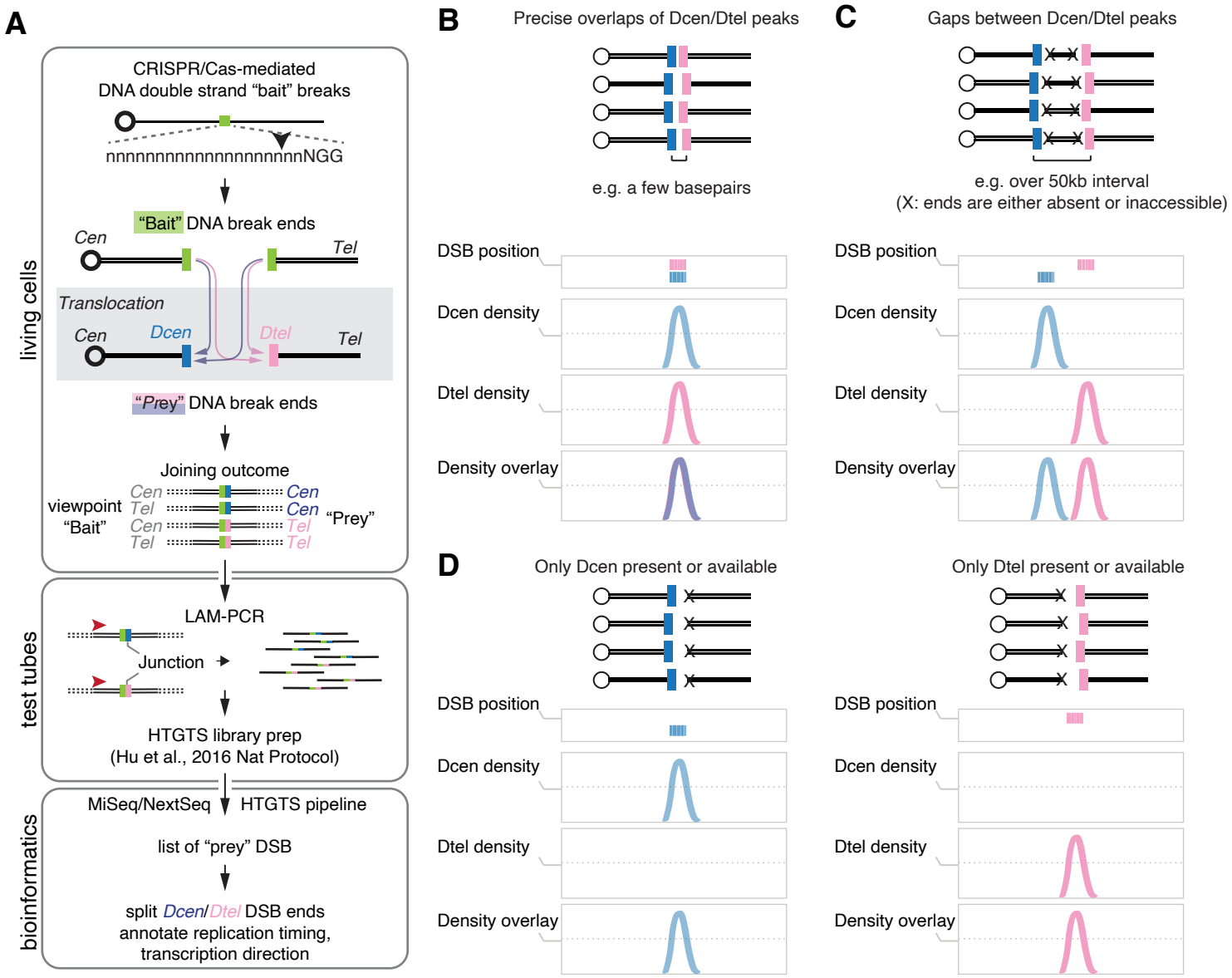

A
living cells
CRISPR/Cas-mediated
DNA double strand "bait" breaks
nnnnnnnnnnnnnnnnnnnNGG
"Bait" DNA break ends
Cen
Tel
Translocation
Cen
Dcen
Dtel
Tel
"Prey" DNA break ends
Joining outcome
viewpoint
"Bait"
Cen
Tel
Cen
Tel
Cen
Cen
Tel
Tel
"Prey"
test tubes
LAM-PCR
Junction
HTGTS library prep
(Hu et al., 2016 Nat Protocol)
bioinformatics
MiSeq/NextSeq
HTGTS pipeline
list of "prey" DSB
split Dcen/Dtel DSB ends
annotate replication timing,
transcription direction
B
Precise overlaps of Dcen/Dtel peaks
e.g. a few basepairs
DSB position
Dcen density
Dtel density
Density overlay
C
Gaps between Dcen/Dtel peaks
e.g. over 50kb interval
(X: ends are either absent or inaccessible)
DSB position
Dcen density
Dtel density
Density overlay
D
Only Dcen present or available
DSB position
Dcen density
Dtel density
Density overlay
Only Dtel present or available
DSB position
Dcen density
Dtel density
Density overlay

Figure S2

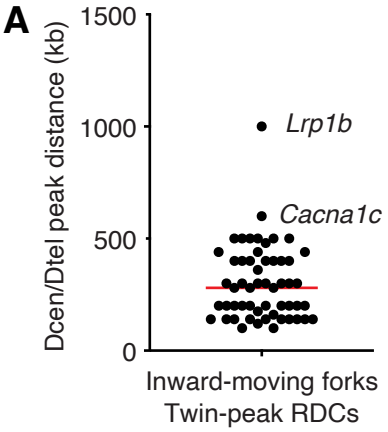

A
1500
1000
500
0
Dcen/Dtel peak distance (kb)
Lrp1b
Cacna1c
Inward-moving forks
Twin-peak RDCs

Figure S3

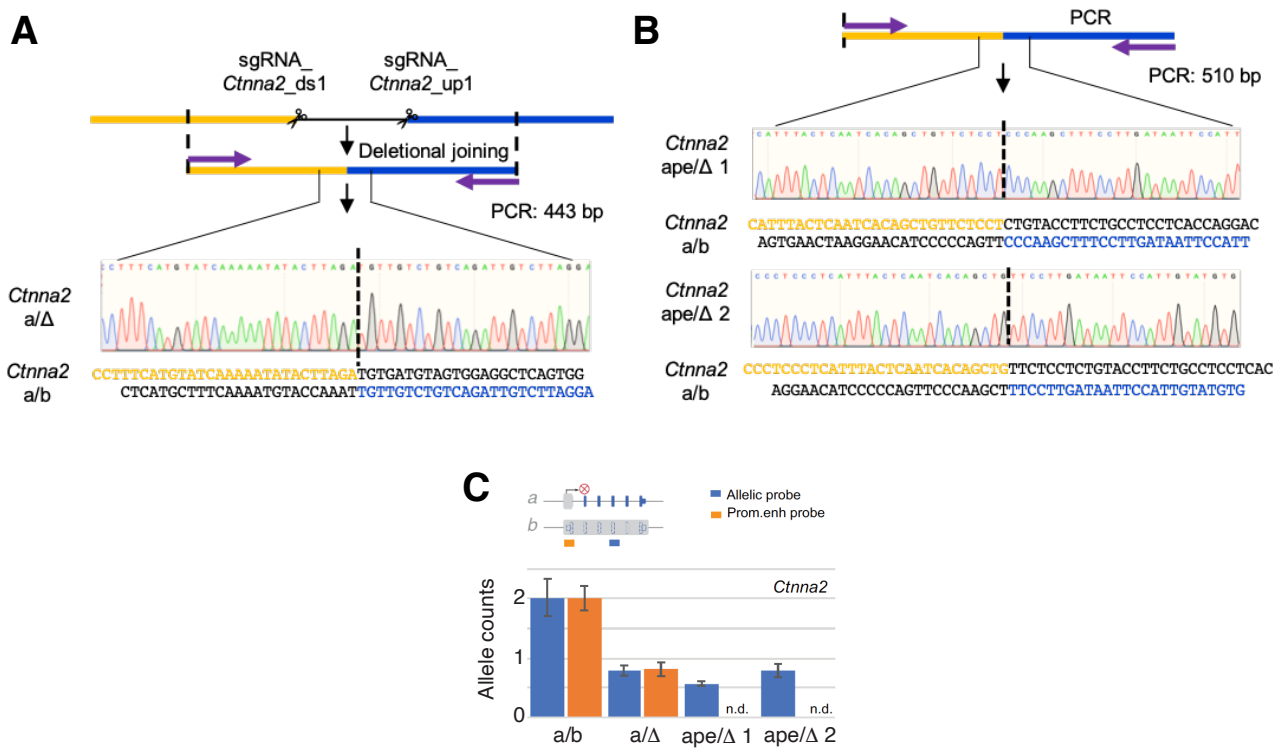

A
sgRNA_
Ctnna2_ds1
sgRNA_
Ctnna2_up1
Deletional joining
PCR: 443 bp
Ctnna2
a/Δ
Ctnna2
a/b
CCTTTCATGTATCAAAAATATACTTAGATGTGATGTAGTGGAGGCTCAGTGG
CTCATGCTTTCAAAATGTACCAAATTGTTGTCTGTCAGATTGTCTTAGGA
B
PCR
PCR: 510 bp
Ctnna2
ape/Δ 1
Ctnna2
a/b
CATTTACTCAATCACAGCTGTTCTCCTCTGTACCTTCTGCCTCCTCACCAGGAC
AGTGAACTAAGGAACATCCCCCAGTTCCCAAGCTTTCCTTGATAATTCCATT
Ctnna2
ape/Δ 2
Ctnna2
a/b
CCCTCCCTCATTTACTCAATCACAGCTGTTCTCCTCTGTACCTTCTGCCTCCTCAC
AGGAACATCCCCCAGTTCCCAAGCTTTCCTTGATAATTCCATTGTATGTG
C
a
b
Allelic probe
Prom.enh probe
Ctnna2
Allele counts
2
1
0
n.d.
n.d.
a/b
a/Δ
ape/Δ 1
ape/Δ 2

Figure S4

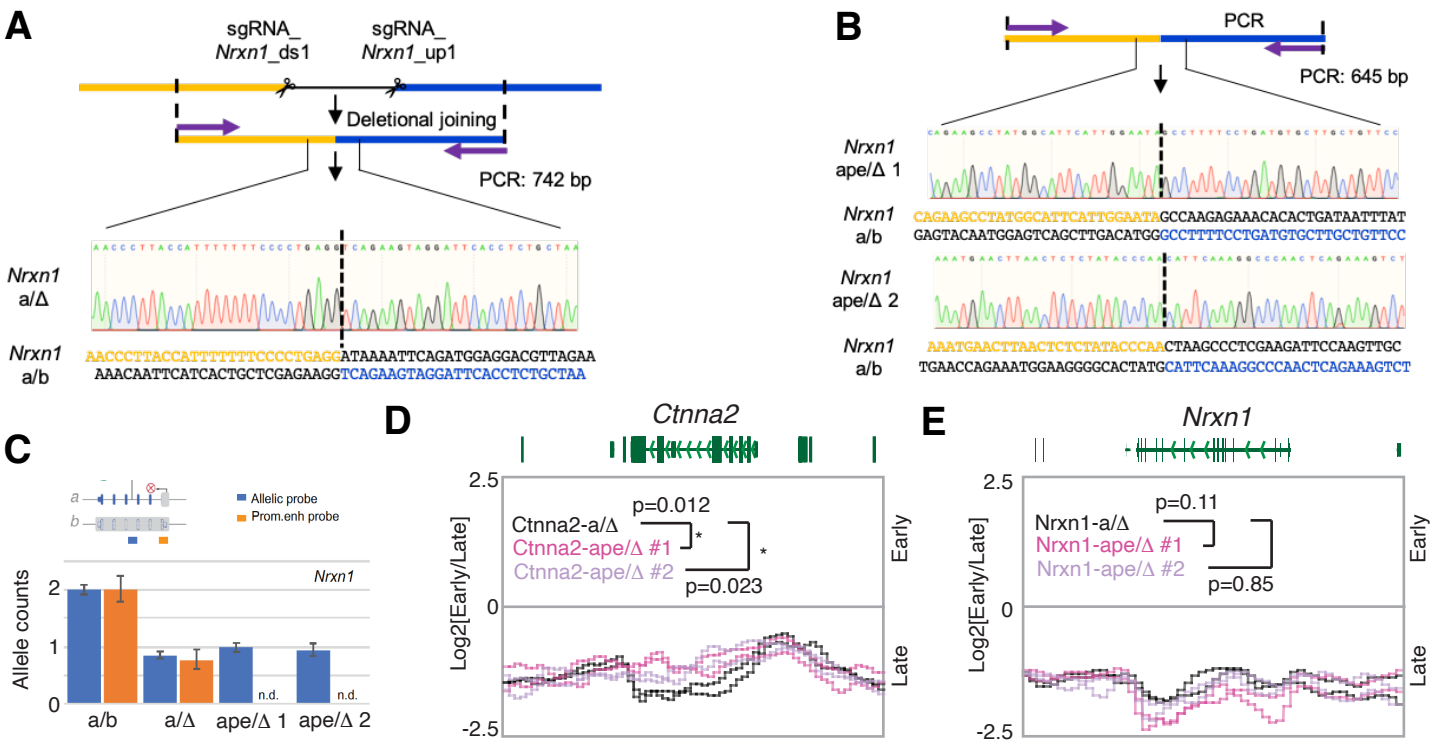

A
sgRNA_ Nrxn1_ds1
sgRNA_ Nrxn1_up1
Deletional joining
PCR: 742 bp
Nrxn1 a/Δ
Nrxn1 a/b
AACCCTTACCATTTTTTTCCCCTGAGCATAAAATTCAGATGGAGGACGTTAGAA
AAACAATTCATCACTGCTCGAGAAGGTCAGAAGTAGGATTCACCTCTGCTAA
B
PCR
PCR: 645 bp
Nrxn1 ape/Δ 1
Nrxn1 a/b
CAGAAGCCTATGGCATTCATTGGAATAGCCAAGAGAAACACACTGATAATTTAT
GAGTACAATGGAGTCAGCTTGACATGGGCCTTTTCCTGATGTGCTTGCTGTTCC
Nrxn1 ape/Δ 2
Nrxn1 a/b
AAATGAACTTAACTCTCTATACCCAACTAAGCCCTCGAAGATTCCAAGTTGC
TGAACCAGAAATGGAAGGGGCACTATGCATTCAAAGGCCCAACTCAGAAAGTCT
C
Allelic probe
Prom.enh probe
Nrxn1
Allele counts
2
1
0
n.d.
n.d.
a/b
a/Δ
ape/Δ 1
ape/Δ 2
D
Ctnna2
2.5
0
-2.5
Log2[Early/Late]
Ctnna2-a/Δ
Ctnna2-ape/Δ #1
Ctnna2-ape/Δ #2
p=0.012
p=0.023
*
*
Early
Late
E
Nrxn1
2.5
0
-2.5
Log2[Early/Late]
Nrxn1-a/Δ
Nrxn1-ape/Δ #1
Nrxn1-ape/Δ #2
p=0.11
p=0.85
Early
Late

Figure S5

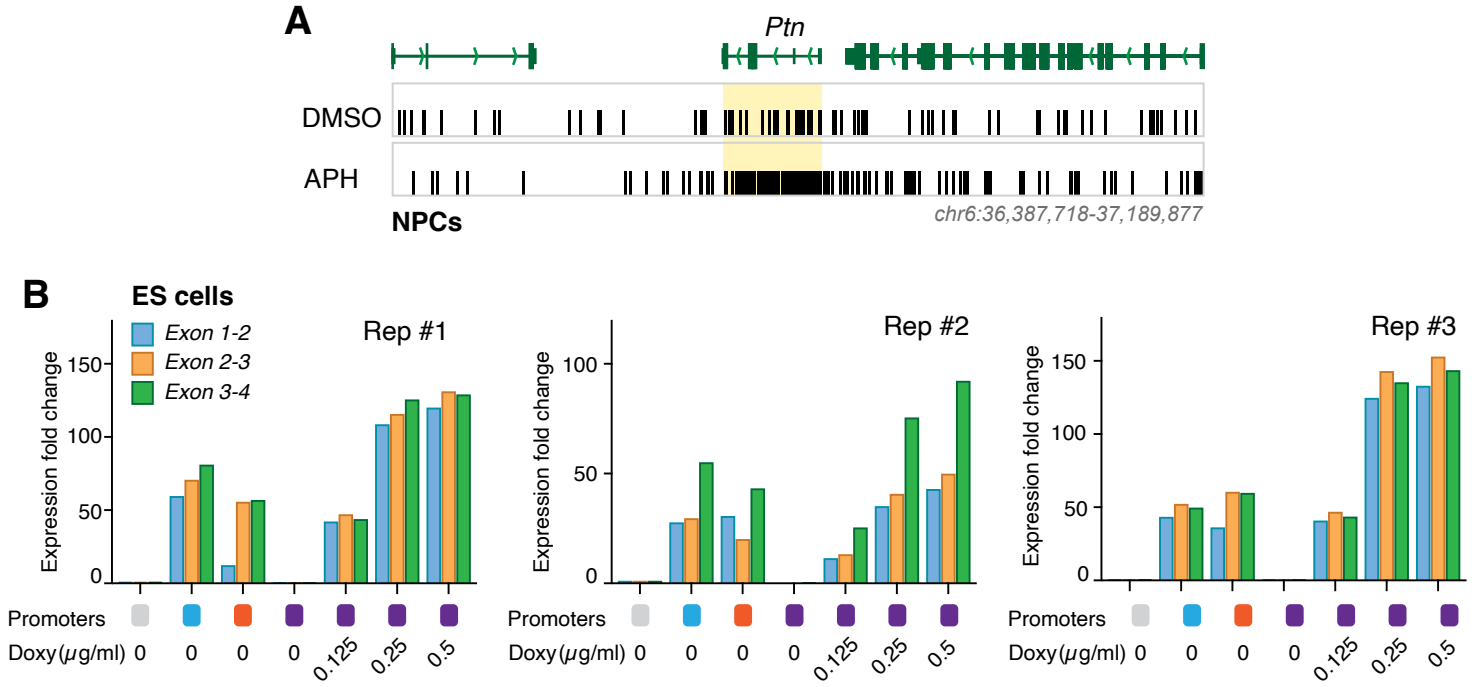
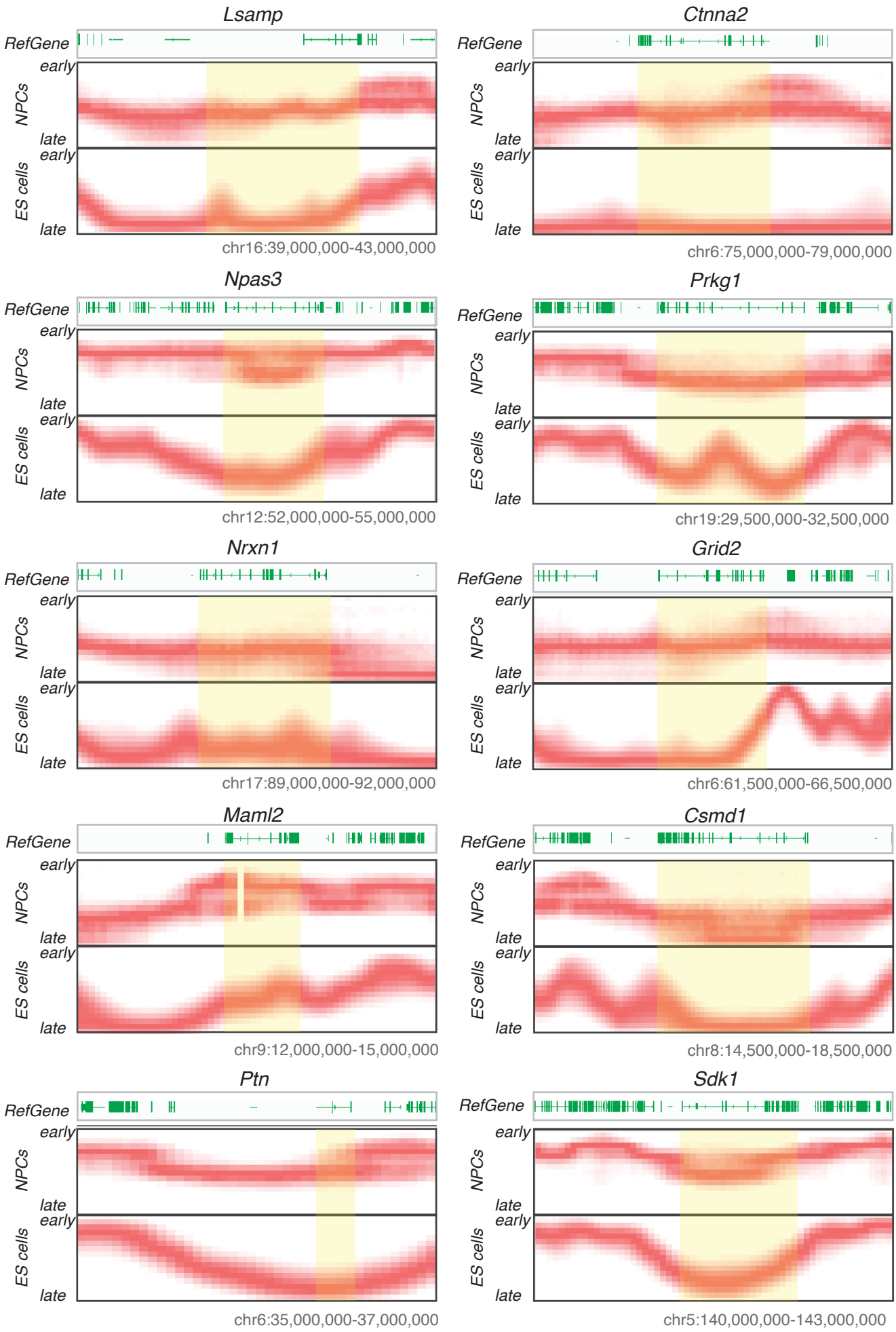

A
Ptn
DMSO
APH
NPCs
chr6:36,387,718-37,189,877
B
ES cells
Exon 1-2
Exon 2-3
Exon 3-4
Rep #1
Rep #2
Rep #3
Expression fold change
Promoters
Doxy (μg/ml)
0
0.125
0.25
0.5

Figure S6



Ionasz and Corazzi et al.