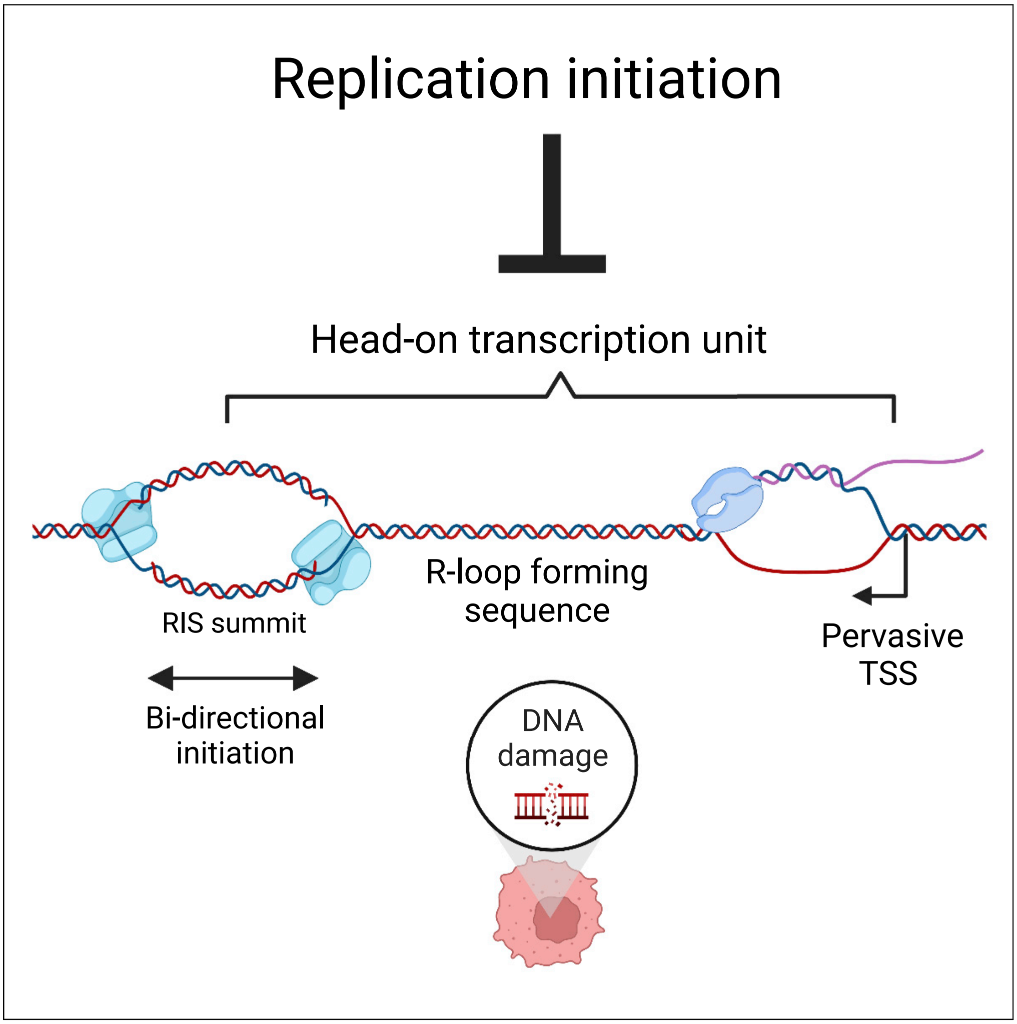
iScience

CellPress OPEN ACCESS

Article

Temporal regulation of head-on transcription at replication initiation sites



Michael Kronenberg, Michael F. Carey

mcarey@mednet.ucla.edu

Highlights

Head-on transcription units occur at replication initiation sites in MCF-7 cells

Head-on transcription units are pervasive and contain R-loop forming sequences

Head-on transcription units are significantly downregulated at the G1/S boundary

Kronenberg & Carey, iScience *26*, 105791
January 20, 2023 © 2022 The Authors.
https://doi.org/10.1016/j.isci.2022.105791

iScience

CellPress OPEN ACCESS

Article

Temporal regulation of head-on transcription at replication initiation sites

Michael Kronenberg[1,2] and Michael F. Carey[1,2,3,4,*]

SUMMARY

Head-on (HO) collisions between the DNA replication machinery and RNA polymerase over R-loop forming sequences (RLFS) are genotoxic, leading to replication fork blockage and DNA breaks. Current models suggest that HO collisions are avoided through replication initiation site (RIS) positioning upstream of active genes, ensuring co-orientation of replication fork movement and genic transcription. However, this model does not account for pervasive transcription, or intragenic RIS. Moreover, pervasive transcription initiation and CG-rich DNA is a feature of RIS, suggesting that HO transcription units (HO TUs) capable of forming R-loops might occur. Through mining phased GRO-seq data, and developing an informatics strategy to stringently identify RIS, we demonstrate that HO TUs containing RLFS occur at RIS in MCF-7 cells, and are downregulated at the G1/S phase boundary. Our analysis reveals a novel spatiotemporal relationship between transcription and replication, and supports the idea that HO collisions are avoided through transcriptional regulatory mechanisms.

INTRODUCTION

DNA replication and transcription are both polymerase-driven reactions that occur on the same DNA template with the potential to spatially and temporally interfere with one another. Prior *in vitro* and *in vivo* studies have shown that head-on collisions between the DNA and RNA polymerases over R-loop forming sequences (RLFS) stably block replication forks through R-loop stabilization on the lagging strand, and G4 quadruplex stabilization on the leading strand.[1–7] In contrast, head-on collisions over non-RLFS, as well as co-directional collisions regardless of sequence environment, are bypassed by the replication fork and do not typically result in DNA breaks.[1,3–5,7] Such findings highlight the need for cells to preserve genome stability by employing mechanisms to avoid head-on collisions over RLFS. However, it is unclear if, where, and at what frequency head-on transcription units (HO TUs) containing RLFS in the template strand occur on the genome.

It is generally assumed that head-on collisions are avoided passively through genome organization. OK-seq, which maps replication fork movement, revealed that replication initiation typically occurs in zones upstream of the transcription start site (TSS) of active genes, and terminates downstream of gene bodies.[8,9] Likewise, optical replication mapping, which maps replication initiation via a single molecule approach, found that most initiation zones (IZs) co-localized with zones identified by OK-seq.[10] The organization suggested by these studies would in principle ensure that leading strand synthesis primarily occurs in a co-directional manner with transcription units. Indeed, recent work evaluating gene transcription and replication fork movement during S-phase found that RNA polymerase at gene transcription start sites (TSS) is bypassed by a co-directional replication fork[11] demonstrating that co-directional collisions are tolerated at these sites.

Although a genome-wide co-directional relationship between replication and transcription would adequately explain how genotoxic collisions are avoided, there are limitations to this model. Although about 2% of the genome is occupied by protein-coding genes, 75–90% of the genome is transcribed.[12,13] Non-coding, or pervasive transcription, occurs both outside gene bodies, as in the form of promoter upstream transcripts (PROMPTs) and enhancer RNAs (eRNAs),[14–17] or inside genes including antisense and sense transcription start site associated transcripts (asTSSa and sTSSa).[16,18,19] Moreover, a subset of RIS localize intragenically across mapping assays.[9,20,21] In tumor cells with activated oncogenes, these intragenic RIS increase in frequency.[22] It is possible that pervasive transcription units, or even genes themselves,

[1]Department of Biological Chemistry, UCLA David Geffen School of Medicine, Los Angeles, CA 90095, USA

[2]Molecular Biology Institute, UCLA, Los Angeles, CA 90024, USA

[3]UCLA Jonsson Comprehensive Cancer Center, 10833 Le Conte Avenue, Los Angeles, CA 90024, USA

[4]Lead contact

*Correspondence: mcarey@mednet.ucla.edu

https://doi.org/10.1016/j.isci.2022.105791

form HO TUs. Of interest, Spt6 depletion was shown to upregulate PROMPTs and eRNAs in HeLa cells, leading to R-loop formation, replication stress and DNA damage.[17] Integrative analysis with a small subset of intergenic replication initiation sites (RIS) showed that PROMPT/eRNA upregulation increased transcription through these sites.[17] Collectively, these data suggest that unscheduled pervasive transcription generates genotoxic head-on collisions.

Several lines of evidence suggest pervasive transcription occurs immediately adjacent to RIS in human cells,[23,24] and indeed might be a functional feature, acting to recruit the replication machinery or distribute the MCM helicase.[23,25,26] However, these studies never assessed transcriptional activity at high resolution relative to RIS locations. If RIS-adjacent transcription converged into the RIS, it would present a source of head-on transcription-replication collisions. Furthermore, work assessing the chromatin and sequence environment at RIS has demonstrated that RIS can contain an adjacent NDR with GC-rich DNA.[27,28] Head-on transcription through this region would presumably potentiate genotoxic collisions.

Collectively, it appears that RIS-adjacent pervasive transcription could generate genotoxic collisions. However, the directionality of transcription at these sites relative to the RIS is currently unclear. Furthermore, the sequence environment of these transcription units is unknown. A key question is how do cells avoid head-on collisions at these locations if RIS-adjacent pervasive transcription occurred at a high frequency over RLFS? In this study, we sought to systematically analyze transcriptional activity near RIS with positional and strand resolution utilizing publicly available datasets generated in the MCF-7 breast cancer cell line. By focusing on transcription within 1 kilobase of a subset of stringently identified RIS, we infer replication fork direction and thus determine the positional relationship between transcription and replication. Surprisingly, we find that pervasive HO TUs rich in RLFS occur frequently at both intergenic and intragenic RIS in asynchronous breast cancer cells. Furthermore, we find that HO TUs are significantly downregulated in cells synchronized at the G1/S phase boundary relative to G0/G1-phase cells, especially at TUs dense in template strand RLFS. Collectively, our study identifies RIS-adjacent pervasive transcription as a source of genotoxic head-on collisions, and implicates the existence of a transcriptional regulatory mechanism that functions to silence this transcription before replisome passage to preserve genome stability.

RESULTS

Identifying high-confidence replication initiation sites in the MCF-7 genome

Does RIS-adjacent pervasive transcription potentiate genotoxic head-on transcription-replication collisions? To investigate this, we decided to leverage publicly available datasets in the MCF-7 breast cancer cell line to map RIS, local transcriptional activity, and the overlapping sequence environment at high resolution. The lack of concordance across RIS mapping technologies[29] and the positional sensitivity needed for downstream analysis led us to first prioritize identifying true RIS loci. A multi-layered approach was employed to identify high confidence RIS (hcRIS) in MCF-7 cells (Figure 1A). We first considered a 'core origin' dataset containing ~65,000 regions with a median size of 700 base pairs, which captured a majority (~80%) of small nascent strand sequencing (SNS-seq) reads across 20 human cell types, and significantly overlapped with pre-RC components and IZs identified by OK-seq.[9,27,28] Approximately ~40% of core origin loci are active in any given cell type.[27] To enrich for RIS in the MCF-7 cell line, we used BEDTools software to intersect core origins with MCF-7 SNS-seq peaks yielding 23,110 loci.[21] To further filter out false positives, we intersected the remaining core origins with an epigenetic signature that predicts binding locations of the origin of replication complex (ORC) with remarkable accuracy.[24] This approach yielded 4,572 hcRIS with a median size of 730 bp.

We validated the identified hcRIS set by assessing their positioning relative to MCF-7 repli-seq replication timing (RT) profiles.[12,23] RT profiles contain an inverted V-apex at sites of replication initiation, and typically apex locations contain one or more bonafide RIS.[23] We reasoned that if hcRIS loci were true positives, then a high percentage should localize within apex regions. Viewing the hcRIS on a browser track with RT data clearly showed positioning at apex locations in the earliest S-phase fraction (G1b) (Figure 1B). To assess whether the hcRIS localized to apexes genome-wide, we assigned an s50 score to each hcRIS. An s50 score was assigned if at least 50% of total RT reads map to a region in a single S-phase fraction. This region is then assigned a label for that fraction, indicating that the region is localizing within an inverted-V apex peaking in the indicated temporal window.[23] 68% of total hcRIS loci were assigned a G1b s50 score as compared to 32% of core origins, 33% of epigenetic signature loci, 26% of SNS-seq peaks, and 8% of randomly selected

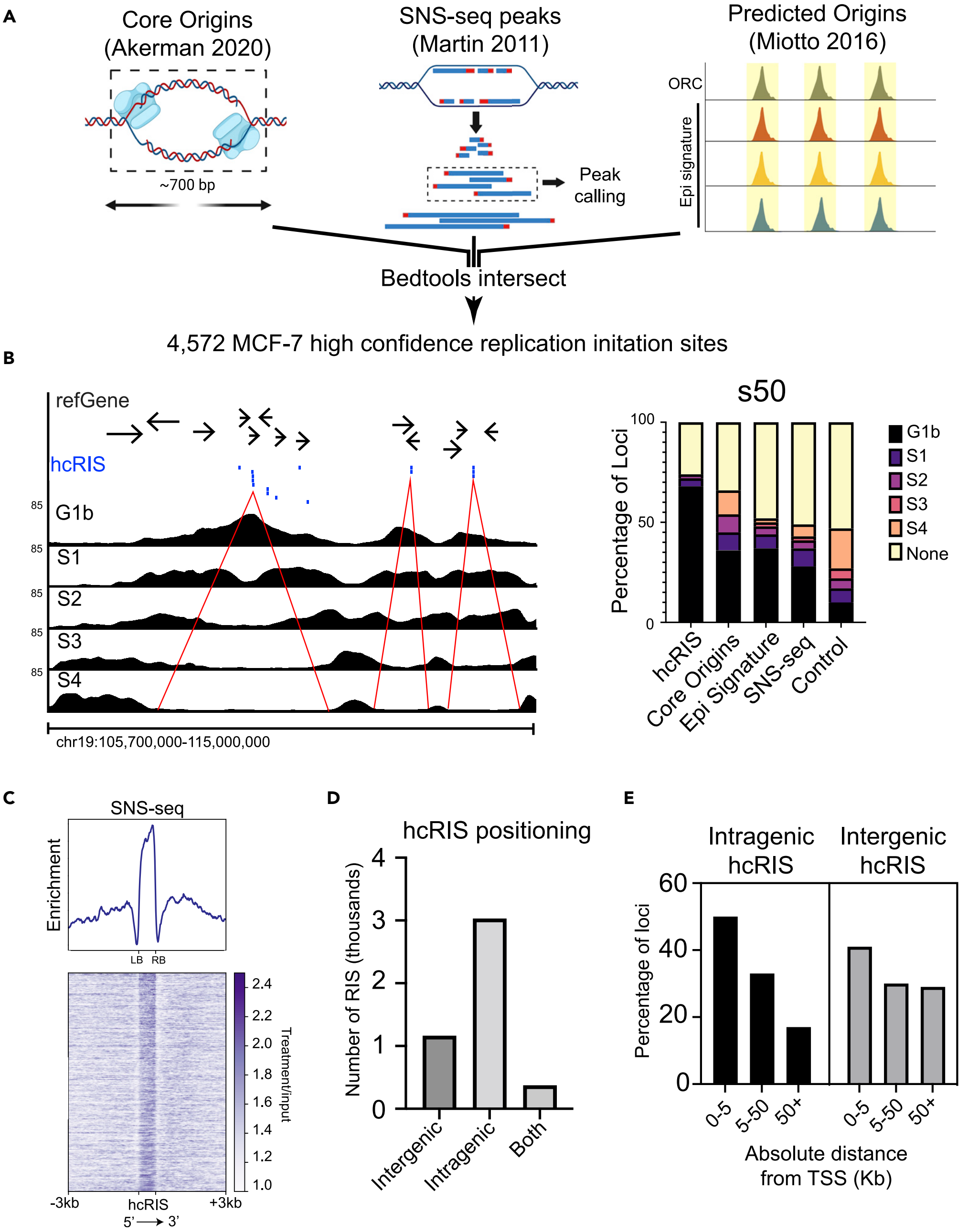


Figure 1. Identifying high confidence replication initiation sites in the MCF-7 genome

(A) Schematic of the strategy used to identify MCF-7 hcRIS.

(B) (Left) Browser track showing hcRIS (top track, blue markers) and RT profiles (Bottom 5 tracks). Tracks are ordered from top to bottom by the earliest S-phase fraction (G1b) to the latest S-phase fraction (S4). Red lines demarcate inverted-V structures. (Right) Distribution of s50 labels across hcRIS, benchmark, and control datasets.

(C) Average profile and heatmap of MCF-7 SNS-seq Poisson enrichment at distance normalized hcRIS loci.

(D) Bar graph showing RIS frequency by position relative to gene bodies.

(E) Bar graphs showing hcRIS frequency by absolute distance relative to the nearest protein-coding TSS.

Dnase-seq peaks, demonstrating that our strategy enriched for true and early replicating RIS (Figure 1B). To further analyze this subset, hcRIS were normalized to the median size of 730 bp and centered on a heatmap encompassing 3 kb upstream and downstream of the left and right boundaries (LB and RB), respectively, based on the Watson strand (Figure 1C). MCF-7 SNS-seq signal within this context reveals a clear

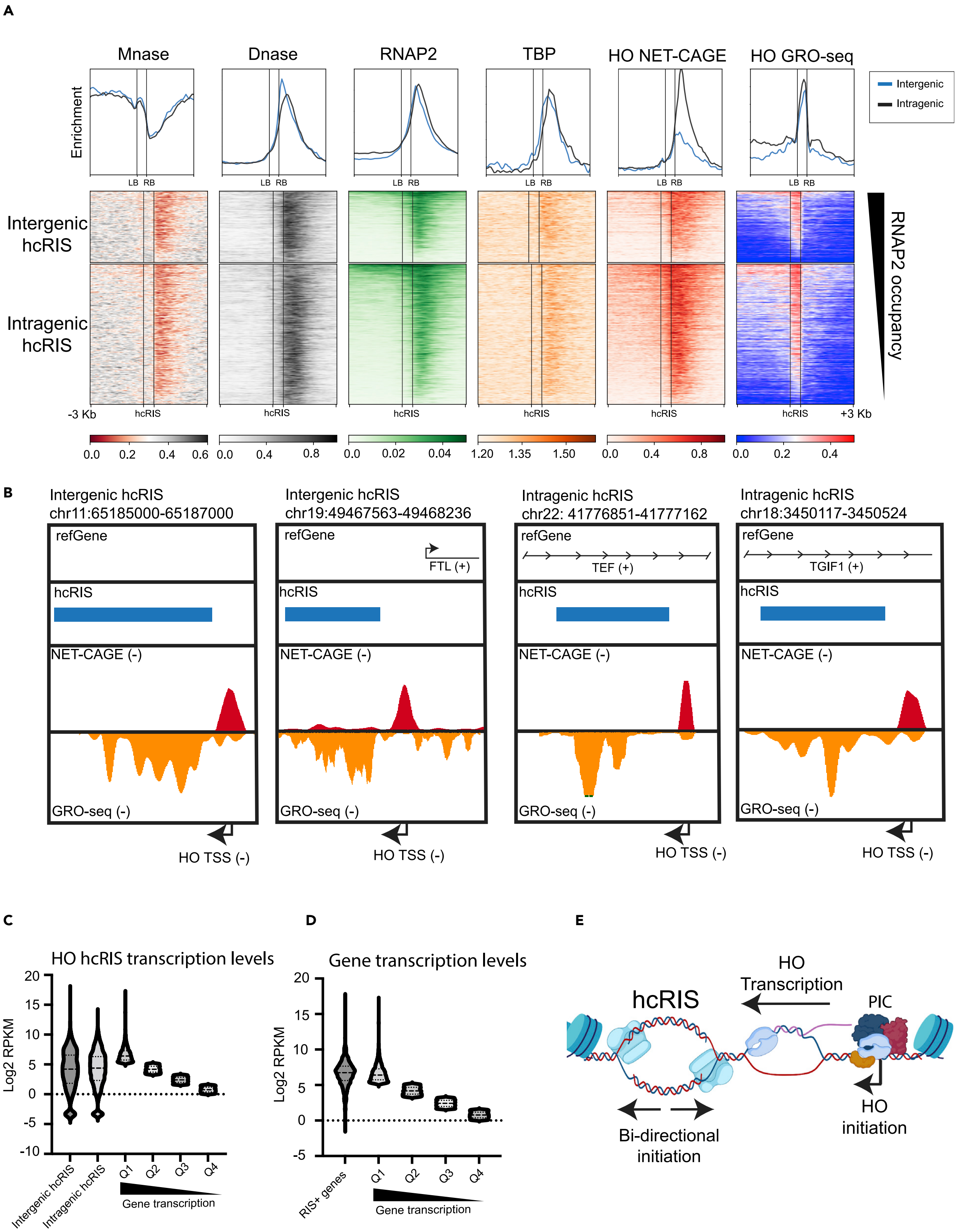

A
Mnase
Dnase
RNAP2
TBP
HO NET-CAGE
HO GRO-seq
Enrichment
LB RB
Intergenic
Intragenic
Intergenic hcRIS
Intragenic hcRIS
RNAP2 occupancy
-3 Kb
hcRIS
+3 Kb
0.0 0.2 0.4 0.6
0.0 0.4 0.8
0.0 0.02 0.04
1.20 1.35 1.50
0.0 0.4 0.8
0.0 0.2 0.4
B
Intergenic hcRIS chr11:65185000-65187000
Intergenic hcRIS chr19:49467563-49468236
Intragenic hcRIS chr22: 41776851-41777162
Intragenic hcRIS chr18:3450117-3450524
refGene
FTL (+)
TEF (+)
TGIF1 (+)
hcRIS
NET-CAGE (-)
GRO-seq (-)
HO TSS (-)
C
HO hcRIS transcription levels
Log2 RPKM
20 15 10 5 0 -5 -10
Intergenic hcRIS
Intragenic hcRIS
Q1 Q2 Q3 Q4
Gene transcription
D
Gene transcription levels
Log2 RPKM
20 15 10 5 0 -5
RIS+ genes
Q1 Q2 Q3 Q4
Gene transcription
E
hcRIS
HO Transcription
PIC
Bi-directional initiation
HO initiation

Figure 2. Head-on transcription occurs at intergenic and intragenic hcRIS

(A) Average profiles and heatmaps of MCF-7 Mnase-seq, Dnase-seq, RNAP2 ChIP-seq, and TBP ChIP-exo Poisson enrichment, and HO NET-CAGE and HO GRO-seq counts per million at distance normalized hcRIS loci. Black lines align with the left and right boundaries (LB and RB) of the hcRIS region.

(B) Browser track examples of HO transcription at intragenic and intergenic hcRIS.

(C) Violin plot showing the distribution of RPKM values for HO GRO-seq reads over subset hcRIS regions, and genic GRO-seq reads over genes split into quartiles by transcription levels.

(D) Violin plot showing the distribution of RPKM values for genic GRO-seq reads over gene bodies containing hcRIS, and genes split into quartiles by transcription levels.

(E) Cartoon model showing positional relationship between HO transcription and replication initiation at hcRIS.

enrichment within the demarcated hcRIS regions (Figure 1C). Among the hcRIS identified, 1,166 localized in intergenic space, 3,030 localized within gene bodies, and 376 spanned gene body termini and adjacent intergenic regions, i.e., both intra and intergenic (Figure 1D), in agreement with the distribution of SNS-seq peaks seen in the MCF-7 cell line.[21] For the intergenic cohort, 40% were within 5 kb of a TSS, 31% were between 5 and 50 kb from a TSS, and 29% were more than 50 kb from a TSS (Figure 1E). For the intragenic cohort, we found that 50%, 34%, and 16% localized in this manner (Figure 1E). With an understanding of the positioning of hcRIS relative to gene units, we next sought to evaluate local transcription at these sites.

Head-on transcription occurs at intergenic and intragenic hcRIS

To assess whether head-on transcription occurs at or near hcRIS and if so, whether it was linked to hcRIS genomic location, we separately evaluated transcription at intergenic and intragenic subsets of hcRIS. We first measured transcription initiation observed within 3 kb of the hcRIS. To positionally map transcription initiation at these sites, we utilized published data from Mnase-seq,[30] Dnase-seq,[12] RNAPII ChIP-seq,[31] and TBP ChIP-exo.[32] In agreement with past findings,[28,33] we find that a strong nucleosome depleted region (NDR) occurs adjacent to hcRIS loci, as indicated by an asymmetric Dnase-seq and Mnase-seq enrichment pattern at hcRIS viewable on a heatmap (Figure 2A, left 2 panels). Of interest, we find that RNAPII ChIP-seq and TBP ChIP-exo signal is enriched within this NDR region across both hcRIS subsets (Figure 2A, middle 2 panels), demonstrating that TBP-containing Pol II preinitiation complexes (PICs) form adjacent to hcRIS. Thus, proximal transcription initiation is a feature of the local hcRIS environment.

To determine if the PICs observed at these sites were initiating transcription ‘head-on’ into hcRIS, we utilized MCF-7 NET-CAGE data, which maps the 5′ ends of nascent RNAs with directional information.[34] We assessed NET-CAGE signal originating at the PIC sites within the downstream NDR and traveling into the upstream hcRIS. We called this head-on (HO) initiation because it marks the start of transcription that converges into an emerging replication fork. Remarkably, we find strong HO initiation signal within the PIC-bound NDR on a global scale (Figure 2A, fifth panel from left), demonstrating that head-on transcripts initiate adjacent to hcRIS.

GRO-seq is a highly sensitive nuclear run-on assay capable of mapping genic and pervasive transcription with strand specificity,[35] including unstable and lowly expressed transcripts. To evaluate transcriptional activity at hcRIS, we utilized asynchronous GRO-seq data generated in the MCF-7 cell line.[36] To assess head-on (HO) transcription, we used the same positional strategy as the NET-CAGE analysis. On a global scale, we found a strong peak of transcription initiating near the border of the hcRIS adjacent to the NDR and peaking within the center of the hcRIS (Figure 2A, rightmost panel). On a local scale, we find clear examples of both HO initiation and transcription at hcRIS loci (Figure 2B). These data reveal that HO transcription is a feature of the local hcRIS environment. HO transcription was evident at both intergenic and intragenic hcRIS (Figures 2A and 2B), demonstrating that it is an intrinsic feature of this hcRIS subset and not due toassociation with gene bodies. Analysis of HO GRO-seq read density at the two subsets of hcRIS relative to gene template strand GRO-seq read density showed that HO transcription occurs at a similar frequency as highly expressed gene transcription (Figure 2C).

Genic transcription could also cause head-on collisions with intragenic RIS. GRO-seq RPKM values for hcRIS-containing genes were similar to that of high to moderately transcribed genes (Figure 2D), suggesting that coding transcription within hcRIS-containing genes occurs at a fairly high frequency across asynchronous cells. Collectively, these data demonstrate that appreciable amounts of transcription occur in the HO orientation in asynchronous MCF-7 cells (Figure 2E).

Pervasive HO TUs containing RLFS are a feature of hcRIS

We next sought to annotate HO TUs at hcRIS to quantify HO transcription frequency, evaluate the template strand sequence at HO transcripts, and assess whether HO transcription was pervasive. We defined HO TUs as regions bookended on one end by an HO NET CAGE-seq peak within 1 kb of an hcRIS border, and on the other the hcRIS summit (Figure 3A). We found that 3,357 (73%) of the 4,572 hcRIS contained at least one HO TU (Figure 3B). In total, we identified 4,567 HO TUs, as multiple units formed at some hcRIS. Viewing NET CAGE-seq and GRO-seq signals at HO TUs on a heatmap clearly demonstrates that HO transcription is initiating at and elongating within the TUs, observably peaking at the hcRIS summit (Figure 3C, left and middle). Thus, in agreement with the analysis in Figure 2, HO transcription is a feature of a majority of hcRIS, and occurs within distinct, identifiable units.

A key question was whether RLFS occur within the template strand of HO TUs, as head-on transcription over RLFS is highly genotoxic when colliding with the replisome.[1] To address this question, we first identified directional RLFS annotated by R-loopDB[37] within the template strand of HO TUs. We found that 76% (3,456/4,567) of HO TUs contain at least one RLFS in their template strand. Mapping RLFS density across HO TUs revealed that template strand RLFS are confined to HO TU bodies, and peak in the HO TU center (Figure 3C, right). A browser track of an individual locus shows clear enrichment of RLFS within the template strand of transcribed HO TUs (Figure 3D). Thus, a majority of HO TUs are transcribed over RLFS, indicating they are likely sources of genotoxic collisions.

Finally, to evaluate whether HO TUs are pervasive in nature, we first evaluated HO TU RPKM values between GRO-seq and RNA-seq datasets from the same study.[36] Pervasive transcripts are typically unstable.[16] RNA-seq, which quantifies steady-state levels of transcription, is not adequate to capture pervasive transcripts because of high levels of turnover. Alternatively, GRO-seq, which captures transcriptional activity via nuclear run-on, is able to effectively quantify pervasive transcripts.[35] If HO TUs are pervasive, they should be underrepresented in RNA-seq relative to GRO-seq data. Indeed, we found that HO TU RPKMs are significantly higher in GRO-seq relative to RNA-seq (Figure 3E). Importantly, active gene RPKMs are underrepresented in GRO-seq relative to RNA-seq, validating our approach (Figure 3E).

We next evaluated whether HO TUs associate with different pervasive transcript species, and if so, at what frequency. We first identified all transcripts belonging to four different pervasive species: promoter upstream transcripts (PROMPTs), enhancer RNAs (eRNAs), antisense TSS-associated RNAs (asTSSa), and sense TSS-associated RNAs (sTSSa) utilizing GRO-seq data[16,38] (Figure S1A). We then categorized HO TUs by whether they overlapped with any of these pervasive transcript classes. We find that 11% of HO TU associations are with PROMPTs, 16% with eRNAs, 18% with asTSSa, 34% with sTSSa, and 21% with transcripts outside these classes (Figure S1B). Finally, we observed GRO-seq and RNA-seq values across HO TUs categorized by transcript class association. We found that HO TUs had higher GRO-seq RPKMs across associations, reinforcing that HO TUs are indeed pervasive in nature (Figure S1C). In aggregate, these analyses support the occurrence of RLFS-rich, pervasive HO TUs at a majority of hcRIS loci (Figure 3F).

Head-on transcription is downregulated at intergenic and intragenic hcRIS at the G1/S boundary

The results from asynchronous MCF-7 cells described above raise the question of how could head-on transcription over RLFS occur at hcRIS without negative effects on cellular fitness? We hypothesized that although head-on transcription occurs at hcRIS during the cell cycle, it might be mitigated before replication initiation. As a majority of the hcRIS in our dataset replicated very early in S-phase (Figures 1C and 1D), we reasoned that head-on transcription at these sites might be downregulated during S-phase entry, at the G1/S boundary. To test this idea, we evaluated HO GRO-seq signal and RPKM distributions at intergenic and intragenic hcRIS between MCF-7 cells synchronized in either G0/G1 phase by hormone starvation, or at the G1/S phase boundary by double thymidine block (Liu et al., 2017). The results show there is a marked decrease in GRO-seq signal in G1/S-phase cells relative to G0/G1-phase cells across both hcRIS subsets (Figures 4A–4C). Importantly, although we observed a small overall decrease in gene transcription between G1/S-phase and G0/G1-phase cells, the magnitude of the HO transcriptional changes at hcRIS are significantly greater, demonstrating that downregulation at the G1/S-phase boundary is biased toward hcRIS (Figure 4D). Moreover, the transcription of genes with proximal upstream hcRIS is not downregulated in S-phase, demonstrating that the effects seen at hcRIS are independent of transcriptional buffering that

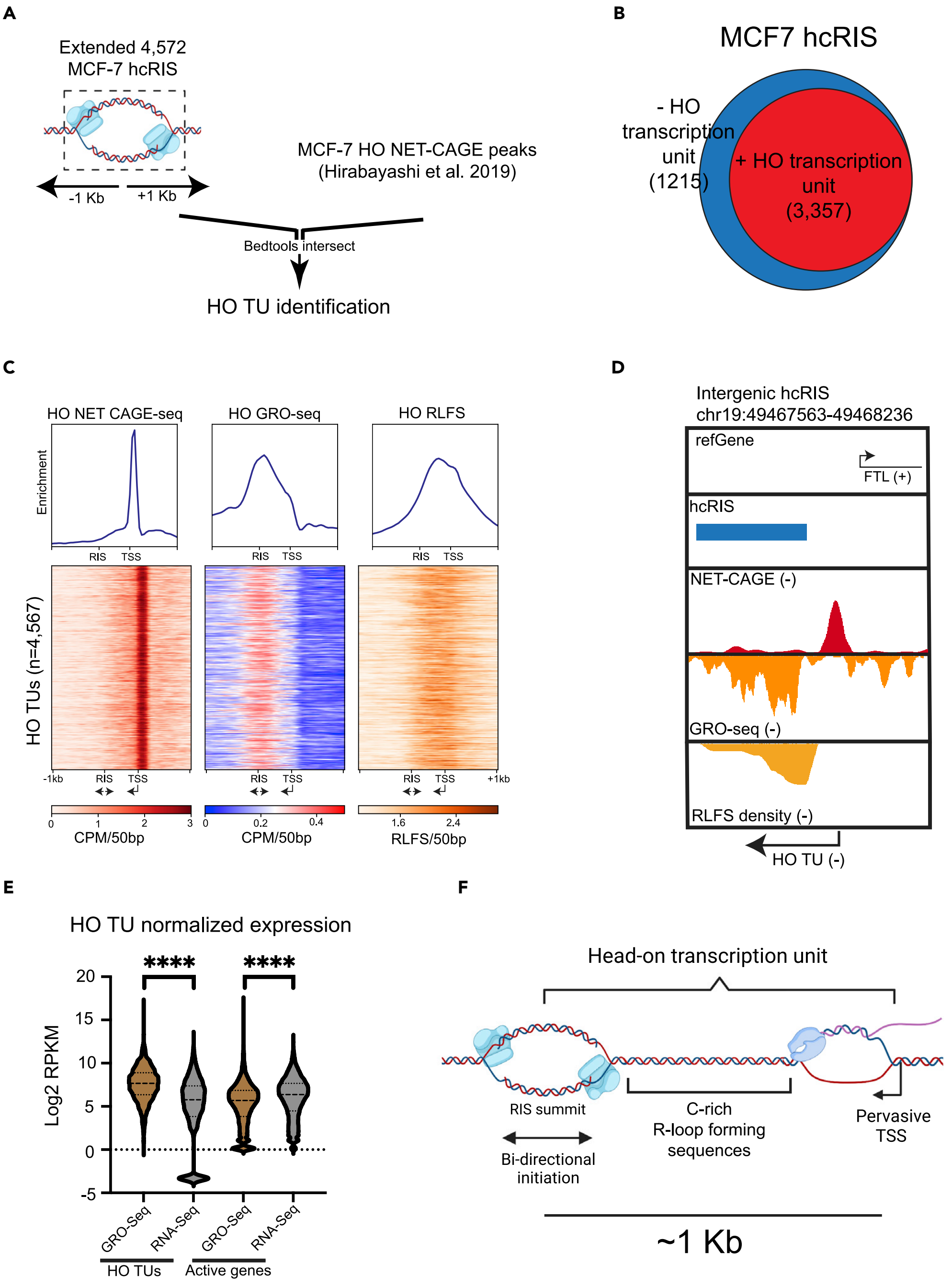


Figure 3. Pervasive HO TUs containing RLFS are a feature of hcRIS

(A) Schematic of strategy to identify HO TUs at hcRIS.

(B) Diagram of total hcRIS demarcated by the presence or absence of at least one HO TU.

(C) Average profiles and heatmaps of HO CAGE-seq/HO GRO-seq (asynchronous) and RLFS density in counts per million and total annotated sequences respectively within 50-bp bins centered on distance normalized HO TUs demarcated by the TSS and RIS summit.

(D) Browser track example of an HO TU.

(E) Violin plot showing the distribution of HO TU or active gene RPKMs from either the GRO-seq or RNA-seq assay ($p < 0.0001$).

(F) Cartoon model of HO TUs with identified features.

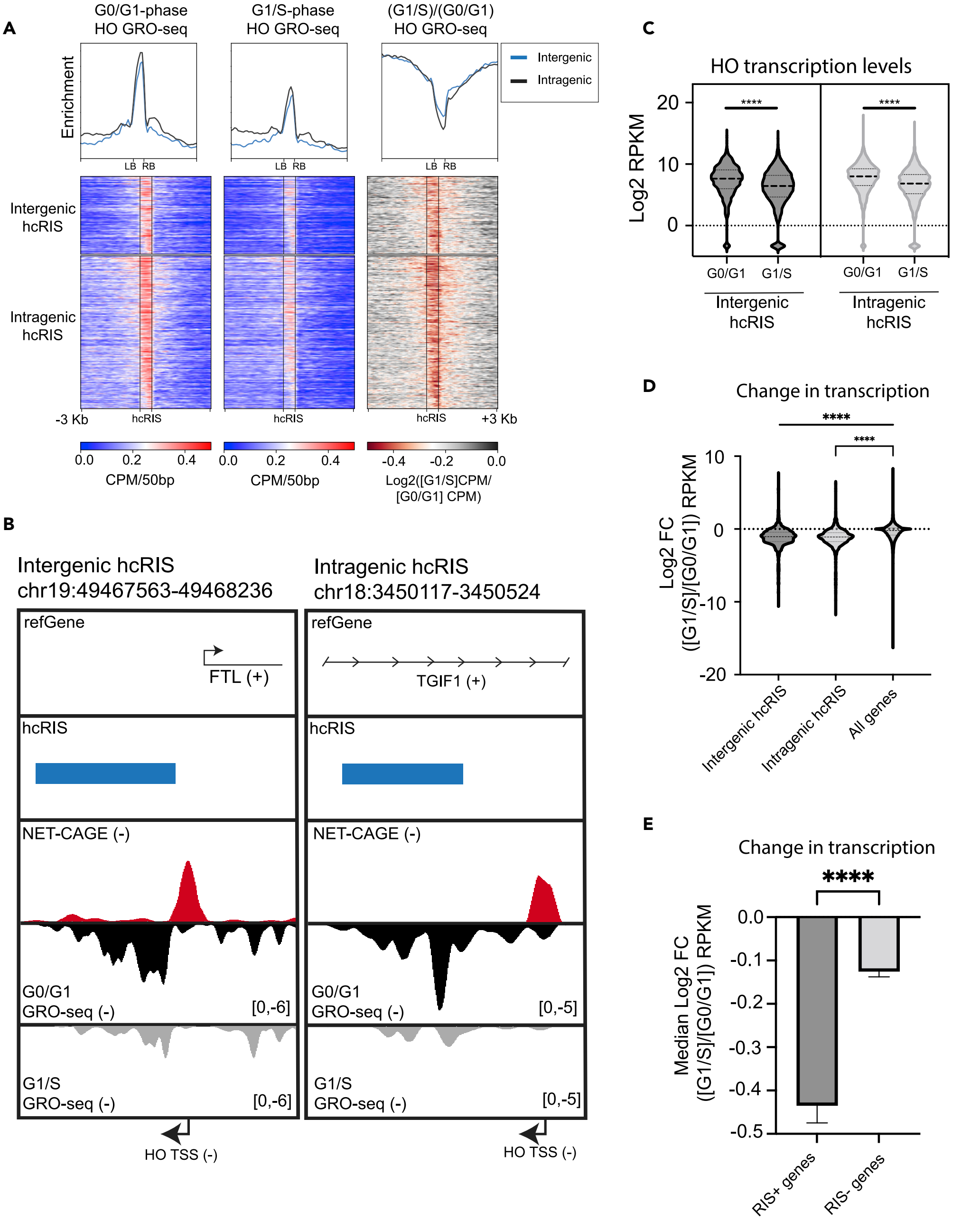


Figure 4. Head-on transcription is downregulated at intergenic and intragenic hcRIS at the G1/S boundary

(A) Average profiles and heatmaps of head-on (HO) GRO-seq signal in counts per million at distance-normalized hcRIS loci (Left and Middle panels). GRO-seq from G0/G1-phase cells (Left). GRO-seq from G1/S-phase cells (Middle). Average profiles and heatmaps of the log2 fold change between G1/S-phase and G0/G1-phase CPM values (Right panel). Black lines align with the left and right boundaries of the hcRIS region.

(B) Browser track examples of changes in HO transcription at intragenic and intergenic RIS.

(C) Violin plot comparing the distribution of RPKM values for GRO-seq reads in the HO orientation across intergenic and intragenic hcRIS regions from G0/G1-phase and G1/S-phase cells (p <0.0001).

(D) Violin plot comparing the distributions of the fold changes in either HO GRO-seq reads or genic GRO-seq reads between G1/S-phase and G0/G1-phase cells across intergenic hcRIS, intragenic hcRIS, and all protein coding genes (p <0.0001).

(E) Bar graph showing the median fold change (95% confidence interval) in protein-coding gene transcription across genes with and without internal hcRIS (p < 0.0001).

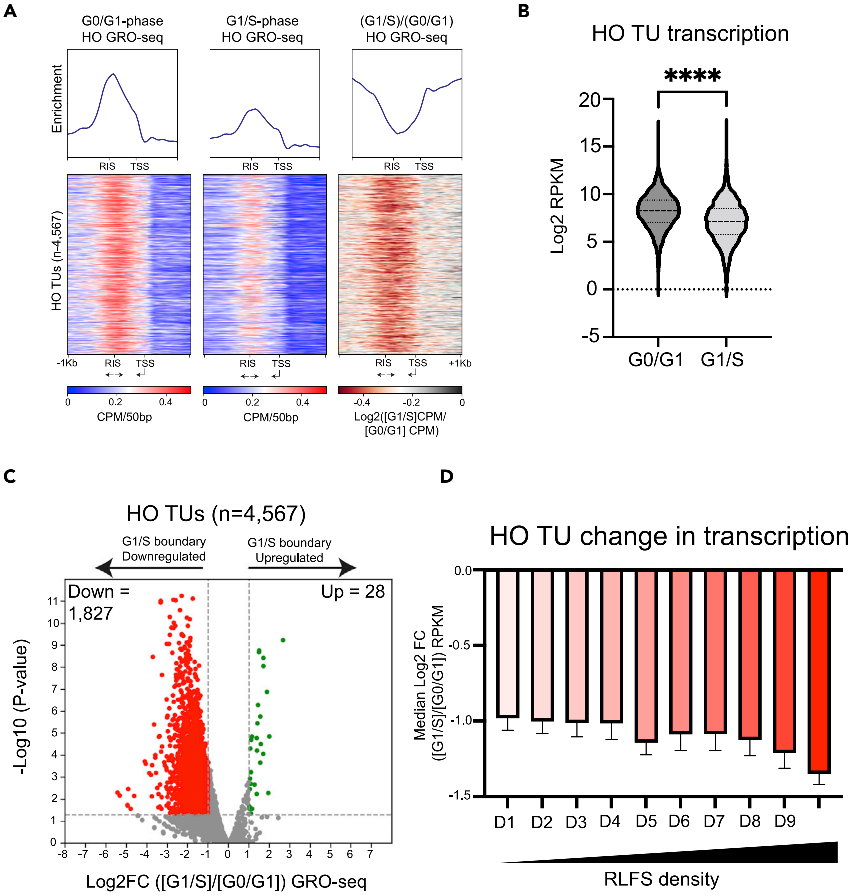


Figure 5. HO TUs are downregulated at the G1/S boundary as a function of RLFS density

(A) Average profiles and heatmaps of head-on (HO) GRO-seq signal in counts per million within 50-bp bins at distance normalized HO TUs (Left and Middle panels). GRO-seq from G0/G1-phase cells (Left). GRO-seq from G1/S-phase cells (Middle). Average profiles and heatmaps of the log2 fold change between G1/S-phase and G0/G1-phase CPM values (Right panel).

(B) Violin plots of HO TU RPKM distributions in G0/G1 and G1/S-phase synchronized cells (p <0.0001).

(C) Volcano plot showing the differential expression of HO TUs between G1/S-phase and G0/G1-phase cells.

(D) Bar graph showing the median log2 fold change (95% confidence interval) between G1/S and G0/G1 HO TU RPKMs across HO TU deciles ordered by increasing RLFS density.

might occur on replicated DNA[39,40] (Figure S2). These analyses indicate that HO pervasive transcription at hcRIS is selectively downregulated during S-phase entry, suggesting that temporally tuned transcriptional regulation at hcRIS might play a role in genome stability.

We also sought to determine if genic transcription through genes containing RISs was downregulated during S-phase entry. Indeed, hcRIS-containing genes are downregulated to a greater degree than genes lacking hcRIS, suggesting that head-on genic transcription is preferentially downregulated during genome replication (Figure 4E). Collectively, the data in Figure 2 through 4 suggest a model in which head-on pervasive and coding transcription occurs during the cell cycle, but is reduced during S-phase entry, potentially to avoid genotoxic collisions with the replisome.

HO TUs are downregulated at the G1/S boundary as a function of RLFS density

We next assessed transcriptional dynamics at HO TUs. In agreement with the previous analysis, we found that HO TU transcription was significantly downregulated in G1/S-phase relative to G0/G1-phase cells (Figures 5A and 5B). Differential expression analysis revealed that 1,827 HO TUs are significantly downregulated in S-phase, whereas only 28 HO TUs are significantly upregulated (Figure 5C). Moreover, significant reductions in HO TU transcription levels at the G1/S boundary are apparent across transcript class associations, demonstrating that suppression upon S-phase entry is a feature of HO TUs independent of pervasive transcript species (Figure S3). Importantly, we found that randomly selected size and transcriptional activity matched TUs within gene bodies did not show G1/S-phase specific downregulation (Figure S4A). Of interest, pervasive transcription units, protein-coding genes, and lincRNAs showed a slight bias toward G1/S-phase downregulation (Figures S4B–S4D). However, comparison of the log2 fold-change distribution across HO TUs and these transcript classes demonstrates that HO TUs experience a significantly greater magnitude of downregulation (Figure S4E). Thus, HO TUs are selectively silenced at the G1/S boundary, and are regulated independently of genic and pervasive transcription broadly speaking.

The decreases we observed in GRO-seq signal in cells enriched at the G1/S boundary relative to G0/G1 synchronized cells could be indicative of transcriptional silencing upon S-phase entry, or alternatively, loss of transcriptionally competent complexes post-collision with an emerging replication fork. For example, collisions could lead to RNAPII stalling, backtracking, or removal. To distinguish between these possibilities, we leveraged phased H3K27ac ChIP-seq data from the same study.[36] H3K27ac is a histone mark that has been shown to stimulate both transcription initiation and elongation, but not form as a consequence of transcription itself.[41,42] Moreover, H3K27ac has been shown to be unaffected by increased RNAP2 stalling at the IgG locus.[43] Thus, downregulation of this mark at the G1/S boundary might indicate loss of a transcriptionally competent environment at HO TUs, independent of post-collision processing. We find that H3K27ac is significantly decreased at HO TUs at the G1/S boundary relative to G0/G1-phase (FiguresS5A and S5B), and that this observed decrease is specific to HO TUs (FiguresS5C), supporting our view that the observed changes in GRO-seq signal are indicative of loss of transcription in a pre-collision setting.

If HO TU silencing is indeed a DNA damage prevention mechanism, we reasoned that the magnitude of silencing would increase with RLFS density, as denser regions would potentiate more genotoxic collisions.[7] Therefore, we analyzed temporal transcriptional changes at HO TUs subset into deciles by increasing RLFS density levels. We observed a clear relationship between increasing RLFS density and G1/S boundary downregulation, suggesting that temporal suppression of HO TUs is likely a mechanism to prevent genotoxic transcription-replication collisions (Figure 5D). In aggregate, we propose that HO TUs potentiate damaging collisions with the replisome, and are actively silenced at the G1/S boundary by still unknown factors before replisome passage to prevent DNA damage.

DISCUSSION

Head-on transcription-replication collisions over RLFS are potent genotoxic events. The co-directional alignment of replication fork movement and gene transcription across the genome is thought to help avoid this type of collision.[9] Our analysis, utilizing multiple published datasets from the MCF-7 breast cancer model, reveals that pervasive HO TUs rich in RLFS form frequently at a subset of RIS in asynchronous cells.[12,21,24,27,32,36] Furthermore, our analysis demonstrates that HO TUs are downregulated at the G1/S boundary, suggesting that collisions are minimized through a temporally tuned transcriptional regulatory mechanism. In support of the idea that head-on transcription is regulated to maintain genome stability, we find that HO TU downregulation correlates with RLFS density. Collectively, our results reveal a surprising spatial relationship between pervasive transcription and replication initiation, and support the presence of a temporally regulated transcriptional axis that functions to prevent DNA damage at a subset of RIS during S-phase.

Recent work assessing transcription and replication across S-phase in immortalized human fibroblasts demonstrated that gene transcription persists during S-phase, and that co-directional replication forks bypass highly transcribed TSS, leaving a gap that is filled by replicative helicases in mitosis.[11] However, given that this study utilized nascent RNA-seq and replication timing profiles to assess transcription and replication activity respectively, it was not adequately powered to evaluate pervasive transcription dynamics at RIS with high resolution. Thus, our work adds to an increasingly complex model of transcription-replication coordination. We propose that co-directional transcription-replication collisions at gene TSS result in replication fork bypass, enabling ‘passive’ avoidance of DNA damage at these sites. However, at RIS themselves, head-on pervasive transcripts over RLFS must be regulated, necessitating ‘active’ aversion of intra-S phase DNA damage at these loci.

Limitations of the study

There are certain limitations to our study. First, our observations were made at a stringently selected subset of RIS that represents about 10% of total RIS on the genome. These RIS show replication timing profiles indicative of replication initiation early in S-phase (Figure 1B). Thus, our observations might not extend to less efficient RIS loci. Indeed, increased transcription has been shown to correlate with replication timing,[23] suggesting that the HO TUs we observe at hcRIS might be specific to loci harboring highly efficient RIS. Second, it is possible that the double thymidine block used to synchronize cells at the G1/S-phase boundary led to the collapse of stalled forks into DNA breaks. It is possible that breaks at hcRIS induced chromatin remodeling events leading to loss of transcription at these sites. However, we see a clear loss of GRO-seq signal starting at the TSS of HO TUs, which localize across a range of distances from the hcRIS summit (Figure 5A). This implies that transcription is being downregulated independent of events at the emerging fork. Furthermore, we find no evidence of a p53-activated transcriptional response (data not shown), suggesting that fork breakage is not occurring in the p53 wild-type MCF-7 cells. However, experimental strategies measuring GRO-seq in cycling cells at optimized timepoints during S-phase entry would be necessary to confirm our assumption.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Cell lines
- METHOD DETAILS
  - ○ Processing of sequencing data
  - ○ hcRIS identification
  - ○ hcRIS validation
  - ○ hcRIS global visualization
  - ○ hcRIS sub-setting by intragenic or intergenic status
  - ○ hcRIS sub-setting by TSS distance
  - ○ Determination of genes with upstream hcRIS
  - ○ Orienting hcRIS to proximal transcription initiation events
  - ○ Head-on transcription unit (HO TU) identification
  - ○ GRO-seq raw data processing
  - ○ Pervasive transcript identification
  - ○ Head-on transcription unit (HO TU) pervasive transcript class association
  - ○ GRO-seq directional heatmap and average profile generation
  - ○ Browser track visualization
  - ○ RPKM calculations
  - ○ RLFS identification and association with HO TUs
  - ○ Differential expression analysis
  - ○ Control TU identification
  - ○ Graphics generation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Statistical tests

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105791.

ACKNOWLEDGMENTS

This work was supported by NIH grant R01 GM074701 to M.F.C., a seed grant from the Jonsson Comprehensive Cancer Center, and a Tobacco Related Disease Research Program 2019B Predoctoral fellowship award T30DT0906 to M.K.

AUTHOR CONTRIBUTIONS

M.K. designed and performed all analyses and wrote the initial draft of the paper. M.F.C. helped write the paper and made conceptual contributions.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 28, 2022
Revised: September 14, 2022
Accepted: December 8, 2022
Published: January 20, 2023

REFERENCES

1. Hamperl, S., Bocek, M.J., Saldivar, J.C., Swigut, T., and Cimprich, K.A. (2017). Transcription-replication conflict orientation modulates R-loop levels and activates distinct DNA damage responses. Cell *170*, 774–786.e19. https://doi.org/10.1016/j.cell.2017.07.043.

2. Lang, K.S., Hall, A.N., Merrikh, C.N., Ragheb, M., Tabakh, H., Pollock, A.J., Woodward, J.J., Dreifus, J.E., and Merrikh, H. (2017). Replication-transcription conflicts generate R-loops that orchestrate bacterial stress survival and pathogenesis. Cell *170*, 787–799.e18. https://doi.org/10.1016/j.cell.2017.07.044.

3. Liu, B., and Alberts, B.M. (1995). Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex. Science *267*, 1131–1137. https://doi.org/10.1126/science.7855590.

4. Mirkin, E.V., and Mirkin, S.M. (2005). Mechanisms of transcription-replication collisions in bacteria. Mol. Cell Biol. *25*, 888–895. https://doi.org/10.1128/MCB.25.3.888-895.2005.

5. Prado, F., and Aguilera, A. (2005). Impairment of replication fork progression mediates RNA polII transcription-associated recombination. EMBO J. *24*, 1267–1276. https://doi.org/10.1038/sj.emboj.7600602.

6. Zardoni, L., Nardini, E., Brambati, A., Lucca, C., Choudhary, R., Loperfido, F., Sabbioneda, S., and Liberi, G. (2021). Elongating RNA polymerase II and RNA:DNA hybrids hinder fork progression and gene expression at sites of head-on replication-transcription collisions. Nucleic Acids Res. *49*, 12769–12784. https://doi.org/10.1093/nar/gkab1146.

7. Brüning, J.G., and Marians, K.J. (2020). Replisome bypass of transcription complexes and R-loops. Nucleic Acids Res. *48*, 10353–10367. https://doi.org/10.1093/nar/gkaa741.

8. Chen, Y.H., Keegan, S., Kahli, M., Tonzi, P., Fenyö, D., Huang, T.T., and Smith, D.J. (2019). Transcription shapes DNA replication initiation and termination in human cells. Nat. Struct. Mol. Biol. *26*, 67–77. https://doi.org/10.1038/s41594-018-0171-0.

9. Petryk, N., Kahli, M., d'Aubenton-Carafa, Y., Jaszczyszyn, Y., Shen, Y., Silvain, M., Thermes, C., Chen, C.L., and Hyrien, O. (2016). Replication landscape of the human genome. Nat. Commun. *7*, 10208. https://doi.org/10.1038/ncomms10208.

10. Wang, W., Klein, K.N., Proesmans, K., Yang, H., Marchal, C., Zhu, X., Borrman, T., Hastie, A., Weng, Z., Bechhoefer, J., et al. (2021). Genome-wide mapping of human DNA replication by optical replication mapping supports a stochastic model of eukaryotic replication. Mol. Cell *81*, 2975–2988.e6. https://doi.org/10.1016/j.molcel.2021.05.024.

11. Wang, J., Rojas, P., Mao, J., Mustè Sadurnì, M., Garnier, O., Xiao, S., Higgs, M.R., Garcia, P., and Saponaro, M. (2021). Persistence of RNA transcription during DNA replication delays duplication of transcription start sites until G2/M. Cell Rep. *34*, 108759. https://doi.org/10.1016/j.celrep.2021.108759.

12. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. https://doi.org/10.1038/nature11247.

13. Hangauer, M.J., Vaughn, I.W., and McManus, M.T. (2013). Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet. *9*, e1003569. https://doi.org/10.1371/journal.pgen.1003569.

14. Berretta, J., and Morillon, A. (2009). Pervasive transcription constitutes a new level of eukaryotic genome regulation. EMBO Rep. *10*, 973–982. https://doi.org/10.1038/embor.2009.181.

15. Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science *322*, 1851–1854. https://doi.org/10.1126/science.1164096.

16. Jacquier, A. (2009). The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat. Rev. Genet. *10*, 833–844. https://doi.org/10.1038/nrg2683.

17. Nojima, T., Tellier, M., Foxwell, J., Ribeiro de Almeida, C., Tan-Wong, S.M., Dhir, S., Dujardin, G., Dhir, A., Murphy, S., and Proudfoot, N.J. (2018). Deregulated expression of mammalian lncRNA through loss of SPT6 induces R-loop formation, replication stress, and cellular senescence. Mol. Cell *72*, 970–984.e7. https://doi.org/10.1016/j.molcel.2018.10.011.

18. McCauley, B.S., and Dang, W. (2022). Loosening chromatin and dysregulated transcription: a perspective on cryptic transcription during mammalian aging. Brief. Funct. Genomics *21*, 56–61. https://doi.org/10.1093/bfgp/elab026.

19. Smolle, M., and Workman, J.L. (2013). Transcription-associated histone modifications and cryptic transcription. Biochim. Biophys. Acta *1829*, 84–97. https://doi.org/10.1016/j.bbagrm.2012.08.008.

20. Langley, A.R., Gräf, S., Smith, J.C., and Krude, T. (2016). Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). Nucleic Acids Res. *44*, 10230–10247. https://doi.org/10.1093/nar/gkw760.

21. Martin, M.M., Ryan, M., Kim, R., Zakas, A.L., Fu, H., Lin, C.M., Reinhold, W.C., Davis, S.R., Bilke, S., Liu, H., et al. (2011). Genome-wide depletion of replication initiation events in highly transcribed regions. Genome Res. *21*, 1822–1832. https://doi.org/10.1101/gr.124644.111.

22. Macheret, M., and Halazonetis, T.D. (2018). Intragenic origins due to short G1 phases underlie oncogene-induced DNA replication stress. Nature *555*, 112–116. https://doi.org/10.1038/nature25507.

23. Dellino, G.I., Cittaro, D., Piccioni, R., Luzi, L., Banfi, S., Segalla, S., Cesaroni, M., Mendoza-Maldonado, R., Giacca, M., and Pelicci, P.G. (2013). Genome-wide mapping of human DNA-replication origins: levels of transcription at ORC1 sites regulate origin selection and replication timing. Genome Res. *23*, 1–11. https://doi.org/10.1101/gr.142331.112.

24. Miotto, B., Ji, Z., and Struhl, K. (2016). Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers. Proc. Natl. Acad. Sci. USA *113*, E4810–E4819. https://doi.org/10.1073/pnas.1609060113.

25. Hoshina, S., Yura, K., Teranishi, H., Kiyasu, N., Tominaga, A., Kadoma, H., Nakatsuka, A., Kunichika, T., Obuse, C., and Waga, S. (2013). Human origin recognition complex binds preferentially to G-quadruplex-preferable RNA and single-stranded DNA. J. Biol. Chem. *288*, 30161–30171. https://doi.org/10.1074/jbc.M113.492504.

26. Gros, J., Kumar, C., Lynch, G., Yadav, T., Whitehouse, I., and Remus, D. (2015). Post-licensing specification of eukaryotic replication origins by facilitated mcm2-7 sliding along DNA. Mol. Cell *60*, 797–807. https://doi.org/10.1016/j.molcel.2015.10.022.

27. Akerman, I., Kasaai, B., Bazarova, A., Sang, P.B., Peiffer, I., Artufel, M., Derelle, R., Smith, G., Rodriguez-Martinez, M., Romano, M., et al. (2020). A predictable conserved DNA base composition signature defines human core DNA replication origins. Nat. Commun. *11*, 4826. https://doi.org/10.1038/s41467-020-18527-0.

28. Foulk, M.S., Urban, J.M., Casella, C., and Gerbi, S.A. (2015). Characterizing and controlling intrinsic biases of lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-quadruplex motifs around a subset of human replication origins. Genome Res. *25*, 725–735. https://doi.org/10.1101/gr.183848.114.

29. Ganier, O., Prorok, P., Akerman, I., and Méchali, M. (2019). Metazoan DNA replication origins. Curr. Opin. Cell Biol. *58*, 134–141. https://doi.org/10.1016/j.ceb.2019.03.003.

30. Shimbo, T., Du, Y., Grimm, S.A., Dhasarathy, A., Mav, D., Shah, R.R., Shi, H., and Wade, P.A. (2013). MBD3 localizes at promoters, gene bodies and enhancers of active genes. PLoS Genet. *9*, e1004028. https://doi.org/10.1371/journal.pgen.1004028.

31. Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773. https://doi.org/10.1093/nar/gky955.

32. Venters, B.J., and Pugh, B.F. (2013). Genomic organization of human transcription initiation complexes. Nature *502*, 53–58. https://doi.org/10.1038/nature12535.

33. Eaton, M.L., Galani, K., Kang, S., Bell, S.P., and MacAlpine, D.M. (2010). Conserved nucleosome positioning defines replication origins. Genes Dev. *24*, 748–753. https://doi.org/10.1101/gad.1913210.

34. Hirabayashi, S., Bhagat, S., Matsuki, Y., Takegami, Y., Uehata, T., Kanemaru, A., Itoh, M., Shirakawa, K., Takaori-Kondo, A., Takeuchi, O., et al. (2019). NET-CAGE characterizes the dynamics and topology of human transcribed cis-regulatory elements. Nat. Genet. *51*, 1369–1379. https://doi.org/10.1038/s41588-019-0485-9.

35. Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–1848. https://doi.org/10.1126/science.1162228.

36. Liu, Y., Chen, S., Wang, S., Soares, F., Fischer, M., Meng, F., Du, Z., Lin, C., Meyer, C., DeCaprio, J.A., et al. (2017). Transcriptional landscape of the human cell cycle. Proc. Natl. Acad. Sci. USA *114*, 3473–3478. https://doi.org/10.1073/pnas.1617636114.

37. Jenjaroenpun, P., Wongsurawat, T., Sutheeworapong, S., and Kuznetsov, V.A. (2017). R-loopDB: a database for R-loop forming sequences (RLFS) and R-loops. Nucleic Acids Res. *45*, D119–D127. https://doi.org/10.1093/nar/gkw1054.

38. Liu, X., Guo, Z., Han, J., Peng, B., Zhang, B., Li, H., Hu, X., David, C.J., and Chen, M. (2022). The PAF1 complex promotes 3′ processing of pervasive transcripts. Cell Rep. *38*, 110519. https://doi.org/10.1016/j.celrep.2022.110519.

39. Padovan-Merhar, O., Nair, G.P., Biaesch, A.G., Mayer, A., Scarfone, S., Foley, S.W., Wu, A.R., Churchman, L.S., Singh, A., and Raj, A. (2015). Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. Mol. Cell *58*, 339–352. https://doi.org/10.1016/j.molcel.2015.03.005.

40. Yunger, S., Kafri, P., Rosenfeld, L., Greenberg, E., Kinor, N., Garini, Y., and Shav-Tal, Y. (2018). S-phase transcriptional buffering quantified on two different promoters. Life Sci. Alliance *1*, e201800086. https://doi.org/10.26508/lsa.201800086.

41. Kang, Y., Kim, Y.W., Kang, J., and Kim, A. (2021). Histone H3K4me1 and H3K27ac play roles in nucleosome eviction and eRNA transcription, respectively, at enhancers. Faseb. J. *35*, e21781. https://doi.org/10.1096/fj.202100488R.

42. Vaid, R., Wen, J., and Mannervik, M. (2020). Release of promoter-proximal paused Pol II in response to histone deacetylase inhibition. Nucleic Acids Res. *48*, 4877–4890. https://doi.org/10.1093/nar/gkaa234.

43. Tarsalainen, A., Maman, Y., Meng, F.L., Kyläniemi, M.K., Soikkeli, A., Budzyńska, P., McDonald, J.J., Šenigl, F., Alt, F.W., Schatz, D.G., and Alinikula, J. (2022). Ig enhancers increase RNA polymerase II stalling at somatic hypermutation target sequences. J. Immunol. *208*, 143–154. https://doi.org/10.4049/jimmunol.2100923.

44. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25. https://doi.org/10.1186/gb-2009-10-3-r25.

45. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

46. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36. https://doi.org/10.1186/gb-2013-14-4-r36.

47. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-seq (MACS). Genome Biol. *9*, R137. https://doi.org/10.1186/gb-2008-9-9-r137.

48. Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187–W191. https://doi.org/10.1093/nar/gku365.

49. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004.

50. Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319. https://doi.org/10.1016/j.cell.2013.03.035.

51. Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S., and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics *26*, 2204–2207. https://doi.org/10.1093/bioinformatics/btq351.

52. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

53. Li, D., Hsu, S., Purushotham, D., Sears, R.L., and Wang, T. (2019). WashU epigenome browser update 2019. Nucleic Acids Res. *47*, W158–W165. https://doi.org/10.1093/nar/gkz348.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Core origin coordinate file	Akerman et al[27]	NCBI Gene Expression Omnibus (GEO): GSE128477
MCF-7 SNS-seq	Martin et al[21]	NCBI Gene Expression Omnibus (GEO): GSE28911
MCF-7 Dnase-seq	John Stamatoyannopoulos, UW	ENCODE: https://doi.org/10.17989/ENCSR000EPH
MCF-7 H3K4me2 ChIP-seq	Bradley Bernstein, Broad	ENCODE: https://doi.org/10.17989/ENCSR875KOJ
MCF-7 H3K27ac ChIP-seq	Bradley Bernstein, Broad	ENCODE: https://doi.org/10.17989/ENCSR752UOD
MCF-7 Repli-seq S1	John Stamatoyannopoulos, UW	ENCODE: https://doi.org/10.17989/ENCSR727ZRP
MCF-7 Repli-seq S2	John Stamatoyannopoulos, UW	ENCODE: https://doi.org/10.17989/ENCSR170QBY
MCF-7 Repli-seq S3	John Stamatoyannopoulos, UW	ENCODE: https://doi.org/10.17989/ENCSR404GFT
MCF-7 Repli-seq S4	John Stamatoyannopoulos, UW	ENCODE: https://doi.org/10.17989/ENCSR831UBH
EJ3 Ini-seq	Langley et al.[20]	European Nucleotide Archive (ENA): PRJEB12207
K562 ORC1 ChIP-seq	Miotto et al.[24]	NCBI Gene Expression Omnibus (GEO): GSE70165
MCF-7 H2A.Z ChIP-seq	Bradley Bernstein, Broad	ENCODE: https://doi.org/10.17989/ENCSR057MWG
MCF-7 RNAP2 ChIP-seq	Vishwanath Iyer, UTA	ENCODE: https://doi.org/10.17989/ENCSR000DMT
MCF-7 TBP ChIP-exo	Venters et al.[32]	NCI read archive: SRA067908
MCF-7 GRO-seq	Liu et al.[36]	NCBI Gene Expression Omnibus (GEO): GSE94479
MCF-7 RNA-seq	Liu et al.[36]	NCBI Gene Expression Omnibus (GEO): GSE94479
MCF-7 H3K27ac ChIP-seq	Liu et al.[36]	NCBI Gene Expression Omnibus (GEO): GSE94479
MCF-7 NET CAGE-seq	Hirabayashi et al.[34]	NCBI Gene Expression Omnibus (GEO): GSE118075
R-loop forming sequences	Jenjaroenpun et al.[37]	http://rloop.bii.a-star.edu.sg/
Experimental models: Cell lines		
MCF-7	ATCC	HTB-22
Software and algorithms		
Bedtools	Quinlan and Hall[44]	https://bedtools.readthedocs.io/en/latest/
Samtools	Li et al., 2009[45]	http://samtools.sourceforge.net/
Tophat2	Kim et al.[46]	https://ccb.jhu.edu/software/tophat/index.shtml
MACS2	Zhang et al.[47]	https://pypi.org/project/MACS2/
Bowtie2	Langmead et al.[44]	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Deeptools	Ramirez et al.[48]	https://deeptools.readthedocs.io/en/develop/
HOMER	Heinz et al.[49]	http://homer.ucsd.edu/homer/
ROSE	Whyte et al.[50]	http://younglab.wi.mit.edu/super_enhancer_code.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Mike Carey (mcarey@mednet.ucla.edu).

Materials availability

This study did not generate new unique reagents

Data and code availability

- This paper analyzes existing, publicly available data. The accession numbers for the datasets are listed in the key resources table.

- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines

Data generated from the MCF-7 cell line (ATCC identifier HTB-22) was used in this study. Culture conditions are described in Liu et al. 2017.[36]

METHOD DETAILS

Processing of sequencing data

All publicly available sequencing datasets used for analysis were downloaded in fastq file format from public repositories, including input files for normalization. All datasets were mapped to the hg19 genome with bowtie2[44] to generate bam alignment files. All bam files were then processed with samtools[45] so that duplicates were removed, and low-quality reads were filtered out. MACS2 peakcall[47] was then used to generate read normalized treatment and background bedgraph files from IP and input controls respectively. MACS2 bdgcmp[47] was then used on normalized IP and input bedgraph files to generate bedgraph files containing genome-wide IP/input Poisson enrichment scores. These bedgraph files were then converted to bigwig files using the bedGraphtoBigWig script from ENCODE[12,51] for downstream analysis using the python deeptools software suite.[48]

hcRIS identification

Core origin summits,[27] MCF-7 SNS-seq peaks,[21] and Dnase-seq peaks[12] localizing within 1 kb of both an MCF-7 H3K27ac ChIP-seq and MCF-7 H3K4me2 ChIP-seq peak were extended 1 kb in each direction using bedtools slop.[52] These extended peaks were then intersected using bedtools intersect.[52] Intersected core origin coordinates were used to represent hcRIS.

hcRIS validation

Samtools bedcov[45] was used to map reads from MCF-7 replication timing datasets (Repli-seq)[12] to hcRIS regions and comparator dataset loci (SNS-seq peaks, ENCODE Dnase-seq peaks, Core origins, and randomly selected Dnase-seq peaks). For SNS-seq peaks, ENCODE Dnase HS peaks, and random Dnase-seq peaks, the center of each peak was extended 1 kb in each direction for mapping using bedtools slop.[52] For hcRIS and core origins, the center of all coordinate locations were taken and extended 1 kb in each direction for mapping using bedtools slop. 4,572 random Dnase-seq peaks were selected through using bedtools shuffle[52] on the Dnase-seq peak dataset and the Linux shell head function. To quantify enrichment at inverted-V apexes of replication timing profiles, normalized repli-seq reads were mapped from all fractions to test regions. If a region contains at least 50% of the total reads from one fraction, then it was marked with an s50 label for that fraction as was done previously.[23]

hcRIS global visualization

For heatmap and average profile generation of hcRIS on a global scale, hcRIS loci were normalized to the same bin number representing the median region size of ∼750 bp and centered within a matrix that also displayed regions 3 kb upstream and downstream of the normalized region (demarcated by a left and right boundary, ‘LB’ and ‘RB’), divided into 50 bp bins using the python deeptools computeMatrix function.[48] The matrix was then sorted by largest to smallest RIS region length using the python deeptools plotHeatmap function.[48] An SNS-seq Poisson enrichment bigwig file was then overlaid onto the matrix via the computeMatrix and plotHeatmap functions.

hcRIS sub-setting by intragenic or intergenic status

Intragenic hcRIS were identified by using bedtools intersect to find hcRIS entirely confined within protein-coding gene body termini as annotated from the GENCODE database.[31] Intergenic hcRIS were identified by using bedtools subtract[52] to identify the remaining hcRIS. If hcRIS both overlapped gene body regions and adjacent intergenic regions, they were categorized as ‘both’ and removed from further analysis.

hcRIS sub-setting by TSS distance

The HOMER annotatePeaks function[49] was used to determine the distance from the nearest protein-coding TSS for each hcRIS location based on the hcRIS center coordinate. hcRIS were then binned by the calculated absolute distance.

Determination of genes with upstream hcRIS

Bedtools intersect was used to find genes with 3 kilobase upstream regions that co-localize with an intergenic RIS. Bedtools intersect was used to filter out all genes that contained internal RIS to generate the final gene set.

Orienting hcRIS to proximal transcription initiation events

hcRIS were uniformly aligned to their proximal NDR region using the python deeptools computeMatrix function[48] and the processed Dnase-seq bigwig file (ENCODE), with the NDR being oriented downstream of the hcRIS region on the matrix. Poisson enrichment scores from the generated TBP ChIP-exo bigwig file and RNAP2 ChIP-seq bigwig file, and counts per million values from stranded NET-CAGE bigwig files oriented convergently relative to the hcRIS were then overlaid onto this aligned matrix using the python deeptools computeMatrix and plotHeatmap functions.[48] All analyses of GRO-seq signal utilize this aligned matrix.

Head-on transcription unit (HO TU) identification

Directional NET CAGE-seq peaks[34] were intersected with regions delimited by a hcRIS center and 1 kb downstream of the hcRIS border proximal to the NDR using bedtools intersect. Minus strand NET CAGE-seq peaks were intersected with hcRIS that formed a downstream NDR, and plus strand NET CAGE-seq peaks were intersected with hcRIS that formed an upstream NDR. Intersected peaks were labeled HO TU TSS, and the cognate hcRIS center point represented the HO TU terminus. Some hcRIS contained multiple HO TUs due to multiple NET CAGE-seq peaks intersecting with the demarcated hcRIS region.

GRO-seq raw data processing

Raw fastq files from[36] were mapped to the hg19 genome with tophat2[46] to produce bam alignment files. Duplicates and low quality reads were removed from bam files via samtools.[45] Replicate bam files were merged for downstream analysis using samtools merge.[45] Merged and QC'd bam files were then converted to stranded bigwig files describing mapped reads in counts per million in python deeptools using the bamCoverage function with the filterRNAstrand option.[48] To generate GRO-seq bigwig files that described asynchronous cell populations, bam files from G1-phase, S-phase, and M-phase MCF-7 cell populations were merged using samtools merge,[45] and converted as previously described. To generate GRO-seq bigwig files for G1-phase and S-phase cell populations, bam files from G1-phase cells and S-phase cells were processed separately.

Pervasive transcript identification

MCF-7 asynchronous GRO-seq datasets were used to perform *de novo* transcript discovery via HOMER software, yielding 82,636 transcripts. Transcripts were labeled as PROMPTs if they were intergenic, within 5 kb upstream of a TSS, and were antisense to the proximal gene. This yielded 5,680 total PROMPTs. Transcripts were labeled as eRNAs if their TSS overlapped with enhancer regions called by the ROSE software with gene TSS exclusion.[50] This yielded 11,564 total eRNAs. Transcripts were labeled as asTSSa if they overlapped with TSS plus 500 bp downstream and were divergent to gene direction. This yielded 6,269 asTSSa. For sTSSa identification, we re-called transcripts from asynchronous GRO-seq data that was filtered to only contain reads 20-90 bp in order to enrich for short pervasive transcripts, yielding 51,492 transcripts. Transcripts were labeled as sTSSa if they overlapped with TSS plus 500 bp downstream and were in the same direction as gene transcription. This yielded 12,276 sTSSa.

Head-on transcription unit (HO TU) pervasive transcript class association

Bedtools intersect was used to find overlap between identified HO TUs and pervasive transcripts by class. Some HO TUs were associated with multiple classes. In these cases, the HO TU was partitioned into both classes for downstream analysis.

GRO-seq directional heatmap and average profile generation

To generate head-on GRO-seq heatmaps and average profiles at hcRIS, GRO-seq stranded bigwig files were directionally mapped to hcRIS loci subset by having either an upstream accessible region or a downstream accessible region based on the plus strand of the genome. Stranded GRO-seq bigwig files were mapped onto the hcRIS matrix as was previously described, using a 150 bp smoothing length.[48] After mapping stranded bigwig files to the directionally subset hcRIS, the matrices were combined via the deeptools computeMatrix Rbind function for visualization of directional GRO-seq signal across all hcRIS.[48]

To observe differences between directional GRO-seq signal at hcRIS between G1 and S-phase cell populations, G1 and S-phase GRO-seq bigwig files were generated from bam files as described above, but a scale factor was applied based on mapped reads from an S2 Drosophila spike-in. Normalized bigwig files could then be mapped as previously described to observe relative signal in counts per million at hcRIS. To assess log2 fold change signal at hcRIS, deeptools bamCompare function was used with the application of a scale factor to produce a bigwig file containing stranded log2 fold change values within 50 bp bins.[48] Bins with values of 0 were replaced with 0.1 for this analysis. These bigwig files could then be directionally mapped onto hcRIS matrices as previously described. The same pipeline was used for heatmap and average profile generation at HO TUs.

Browser track visualization

Bigwig files generated as previously described were directly visualized in the web-based WashU genome browser.[53]

RPKM calculations

Merged and QC'd bam files generated from fastq files from (Liu et al. 2017)[36] as previously described were converted to sam files, separated by strand, reconverted to bam files, and indexed using samtools.[45] To find HO RPKMs, samtools bedcov was used to map reads from stranded bam files directionally onto hcRIS regions subset by location of the accessible region on the plus strand. Subsequent files containing head-on mapped read information for each hcRIS subset were then concatenated. Mapped reads within hcRIS regions were then normalized per kilobase as well as per million mapped reads to give an RPKM value. All RPKM values were log2 transformed for distribution analysis and statistical tests. A similar workflow was used to calculate gene RPKMs, using gene body regions separated by strand to map reads to the template strand via samtools bedcov. For gene quartile separation, genes were filtered out if RPKM <1. Remaining genes were then separated into quartiles based on RPKM values (Q1>Q2>Q3>Q4) for analysis. All violin plot RPKM visualizations were generated via PRISM 9 statistical software.

RLFS identification and association with HO TUs

R-loopDB (http://rloop.bii.a-star.edu.sg/) is an online database containing coordinate files for bioinformatically predicted R-loop forming sequences across model genomes.[37] The merged RLFS coordinate file for the hg19 genome was downloaded and separated by strand. To identify HO TUs that contained RLFS in the head-on transcription template strand, bedtools intersect was used to find HO TUs that overlapped with the directionally appropriate stranded RLFS coordinates.

To generate an RLFS heatmap and average profile at HO TUs, a bedgraph file describing RLFS frequency per 50 bp bin across the hg19 genome was generated via IGB. This file was converted to a bigwig file as previously described and used as an input for python deeptools analysis.

To rank-order hcRIS by RLFS density, a bedgraph file describing RLFS frequency per 50 bp bin across the hg19 genome was generated via IGB and converted to a bigwig file as previously described. This file was used as an input along with hcRIS coordinates for python deeptools analysis. Output files describing RLFS density within individual hcRIS units were rank-ordered and separated into deciles.

Differential expression analysis

Tag directories from G1 and S-phase GRO-seq replicate bam files were generated via HOMER software. A raw read count table was then generated using the HOMER analyzeRepeats script describing the reads mapping from these files to a designated gtf file describing genomic locations of interest. This table was then used as an input for the HOMER getDiffExpression script, which utilizes DESeq2 to generate a file describing Log2 fold change and P-value between conditions at each location of interest. The resulting file was then used as input to be processed by the bioinfokit python program to produce a volcano plot. Predetermined thresholds for significance were less than or equal to a p value of 0.05 and a log2 fold change of 1 or -1.

Control TU identification

Bedtools random was used to generate a bed file of random genomic locations at the median size of HO TUs (760 bp). Genes were then filtered so that only 'active' genes, denoted as the 10,000 most highly expressed genes, were considered. The random loci bed file was then intersected with active gene bodies to produce a bed file describing random HO TU sized regions within actively transcribed genes. 4,567 TUs were then randomly selected to be a representative dataset for downstream analysis.

Graphics generation

All visual graphics in manuscript were created with BioRender.com.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical tests

P-values generated from either RPKM, Log2 fold change, or total read distribution comparisons were calculated using the unpaired parametric T-test in Prism GraphPad. Statistical details of the experiments can be found in the figure legends.