Title
Deciphering and targeting transcription-replication coordination in cancer

Author
Kronenberg, Michael David

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Deciphering and targeting transcription-replication coordination in cancer

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Molecular Biology

by

Michael David Kronenberg

2022

ABSTRACT OF THE DISSERTATION

Deciphering and targeting transcription-replication coordination in cancer

by

Michael David Kronenberg

Doctor of Philosophy in Molecular Biology

University of California, Los Angeles, 2022

Professor Michael Carey, Chair

Head-on collisions between the replication and transcription machinery over R-loop forming sequences potently stall replication forks. Stalled fork structures can then be converted into DNA breaks, leading to apoptosis or growth arrest of cycling cells. The mechanisms which coordinate transcription and replication to avoid these genotoxic collisions are therefore critical for cellular fitness, especially in the case of rapidly dividing tumor cells. However it is unclear if coordination occurs passively through globally encoded co-directionality between transcription units and replication forks, or actively, through transcriptional regulatory mechanisms that function to silence head-on transcripts during S-phase. ‘Active’ coordination would imply that transcriptional regulators could be effectively targeted in cancer to induce collisions and subsequent tumor cell killing. However, the ‘active’ coordination model has never been systematically assessed. In this dissertation, we present work demonstrating that head-on transcription over R-loop forming sequences occurs at a high frequency during the

cell cycle across tumor cell types, that this transcription is temporally downregulated during S-phase, and that INO80 and MOT1 are leveraged in NSCLC to suppress genotoxic TRCs and preserve tumor cell viability. These results suggest that transcriptional regulation is imperative to genome stability, and transcriptional regulators serve as promising targets for the treatment of NSCLC.

The dissertation of Michael David Kronenberg is approved.

Hilary Ann Coller

Douglas L. Black

Heather R. Christofk

Alexander Hoffman

Michael F. Carey, Committee Chair

University of California, Los Angeles

2022

For my Grandparents

TABLE OF CONTENTS

ABSTRACT OF THE DISSERTATION……ii

COMMITTEE PAGE……iv

DEDICATION……v

TABLE OF CONTENTS……vi

LIST OF FIGURES……ix

ACKNOWLEDGEMENTS……xi

VITA……xii

Chapter 1: Introduction……1

Overview……2

Replication Stress……3

Transcription-replication collisions……5

Transcription-replication coordination……8

Induction of genotoxic TRCs as a strategy to treat cancer……11

INO80 and MOT1……12

Non-Small Cell lung cancer……14

Approach, significance, and summary……15

Figures……18

Figure legends……25

References……28

Chapter 2: Temporal regulation of head-on transcription at replication initiation sites……37

Abstract……38

Introduction..……..……...38

Results……42

Discussion……..51

Acknowledgements……53

Figures……54

Figure legends……...58

Supplemental figures……..…61

Supplemental figure legends……..……67

Materials and methods……70

References……83

Chapter 3: INO80 and MOT1 regulate head-on transcription units in Non-Small Cell Lung cancer……91

Abstract……..……92

Introduction...……....93

Results……..….96

Discussion……..105

Acknowledgements……..……107

Figures……..……108

Figure legends……..……112

Supplemental figures……..……..……114

Supplemental figure legends……....118

Materials and methods……..…120

References……..….133

Chapter 4: Mapping cisplatin-induced DNA damage in Non-Small Cell Lung cancer……..…141

Abstract……142

Introduction..……....143

Results……..…144

Discussion…….……146

Acknowledgements…….…..…147

Figures……..…148

Figure legends……..150

Materials and methods……151

References……..…..154

Chapter 5: Conclusions……..157

Transcription-replication coordination……..……158

INO80 and MOT1 in NSCLC……..……160

References……163

LIST OF FIGURES

Figure 1.1……..18

Figure 1.2……..19

Figure 1.3……..20

Figure 1.4……..21

Figure 1.5……..22

Figure 1.6……..23

Figure 1.7……..24

Figure 2.1……..……54

Figure 2.2……..…55

Figure 2.3……..…56

Figure 2.4……..…57

Supplemental figure 2.1……61

Supplemental figure 2.2……62

Supplemental figure 2.3……....63

Supplemental figure 2.4……....64

Supplemental figure 2.5……...65

Supplemental figure 2.6……....66

Figure 3.1……108

Figure 3.2……109

Figure 3.3……110

Figure 3.4……111

Supplemental figure 3.1……114

Supplemental figure 3.2……..…..115

Supplemental figure 3.3……...116

Supplemental figure 3.4……..…..117

Figure 4.1……148

Figure 4.2……149

ACKNOWLEDGEMENTS

I want to first thank my family, who have made this journey possible. My grandparents, Dave and Ortencia, and Jack and Marylin, my parents, Karl and Nan, and my brother, Brian. They are the wind beneath my wings. I want to thank the numerous friends I've made at UCLA who have helped me in the good times and bad, both in and out of the lab. I want to thank my colleagues that I had the pleasure of working with in the Carey lab, including Fei Sun, Xianglong Tan, Yong Xue, Biviana Lie, Kevin Avelar, and Collette Araque. My interactions with them through the years has helped shape me as a scientist. I'd also like to thank my committee members who have guided me on my PhD journey; Heather Christofk, Hilary Coller, Alexander Hoffman, and Doug Black. Finally, I'd like to thank my advisor, Mike Carey, who has always supported me through the trials and tribulations of bench work. His patience and mentorship has been key to my professional development.

VITA

Education and professional

2016 Bachelor of Arts in Biology with Honors, Boston College; Chestnut Hill, CA

2016-present Molecular Biology Institute Doctoral Program, Gene Regulation Home Area; University of California, Los Angeles; Los Angeles, CA

2017-2018 UCLA Cellular and Molecular Biology Training Program T32 Fellowship Award; University of California, Los Angeles; Los Angeles, CA

2019-2022 Tobacco-Related Disease Research Program Pre-Doctoral Fellowship Award; University of California, Los Angeles; Los Angeles, CA

Publications

Kronenberg M, Carey M. Temporal Regulation of Head-on Transcription at Replication Initiation Sites. Cell Reports. “in review”

Sun F, Chronis C, **Kronenberg M**, Chen XF, Su T, Lay FD, Plath K, Kurdistani SK, Carey MF. Promoter-enhancer communication occurs primarily within insulated neighborhoods. Molecular cell. 2019 Jan 17;73(2):250-63.

Sun F, Sun T, **Kronenberg M**, Tan X, Huang C, Carey MF. The Pol II preinitiation complex (PIC) influences Mediator binding but not promoter–enhancer looping. Genes & development. 2021 Aug 1;35(15-16):1175-89.

Presentations

American Society of Biochemistry Transcriptional Regulation Chromatin and RNA Polymerase II Special Symposium. October 2018. **Kronenberg M**, Carey M. Intergenic pervasive transcription is globally upregulated in Non-Small Cell Lung Cancer and potentiates transcription-replication conflicts. (Poster)

Cold Springs Harbor Laboratory, Mechanisms of Eukaryotic Transcription conference. August 2019. **Kronenberg M**, Carey M. INO80 prevents DNA damage and represses pervasive transcription near replication initiation sites in Non-Small Cell Lung Cancer. (Poster)

Jonsson Comprehensive Cancer Center Annual Retreat. May 2022. **Kronenberg M**, Carey M. Pervasive transcription constitutes a threat to cancer genome stability and loss of control is linked to DNA damage induction in Non-Small Cell Lung cancer**.** (Poster)

Chapter 1: Introduction

Overview

Head-on transcription-replication collisions (HO TRCs) over R-loop forming sequences (RLFS) (hereby referred to as genotoxic TRCs), potently stall replication forks and generate DNA breaks in in-vitro and episomal systems (Bruning and Marians, 2020; Hamperl et al., 2017; Helmrich et al., 2013; Kumar et al., 2021; Prado and Aguilera, 2005). The high fitness cost of genotoxic TRCs suggests that dividing cells must suppress these events to maintain viability. However, it is unclear how genotoxic TRCs are avoided within the transcriptionally active landscape of the genome. A 'passive' model proposes that genotoxic TRCs are avoided through encoded co-directionality between transcription units (TUs) and replication forks (Chen et al., 2019; Petryk et al., 2016; Wang et al., 2021). Alternatively, an 'active' model proposes that head-on transcription over RLFS occurs frequently during the cell cycle but is silenced during S-phase to mitigate collisions. The 'active' model is attractive, as it implies that transcriptional regulators could be therapeutically targeted to induce the killing of rapidly dividing tumor cells. *However, it is unclear how frequently head-on transcription over RLFS occurs on the genome, whether this transcription is temporally silenced, and if so, what the regulators of such TUs might be*. **This work addresses these knowledge gaps, with the two-fold goal of testing the hypothesis that the active model reflects transcription-replication coordination in tumor cells, and ascertaining whether the transcriptional regulators INO80 and MOT1 function to prevent genotoxic TRCs in Non-Small Cell Lung cancer (NSCLC).** This introduction will detail the consequences of replication fork stalling, illustrate the effects of different kinds of transcription-replication collisions on replication fork processivity, discuss evidence for

both the passive and active model of transcription-replication coordination, rationalize targeting transcription-replication collisions in cancer for therapeutic purposes, and present logic for investigating the transcriptional regulators INO80 and MOT1 as candidate suppressors of genotoxic TRCs in NSCLC.

Replication stress

DNA replication is a highly organized, stepwise process. In G1-phase cells, ~50,000 replication origins, or replication initiation sites (RIS), are set through recruitment of the 6-subunit origin of replication complex (ORC) to accessible regions enriched for G-quadruplex (G4) secondary structures (Akerman et al., 2020; Dellino et al., 2013; Foulk et al., 2015; Hoshina et al., 2013; Kumagai and Dunphy, 2020; Langley et al., 2016). ORC binding initiates a recruitment cascade, leading to the binding of two MCM2-7 helicase hexamers, thus 'licensing' the origin for downstream activation (Fragkos et al., 2015; Ganier et al., 2019). Throughout S-phase, ~20% of licensed origins are phosphorylated via transient DDK4 and CDK2 activity, leading to association of CDC45 and the GINS complex with the MCM2-7 helicase, and subsequent formation of the activated CMG complex. CMG then catalyzes DNA duplex unwinding, and the bi-directional firing of competent replisomes (Fragkos *et al.*, 2015). Processive replisomes consist of the CMG helicase complex on the leading edge, followed by replicative DNA polymerases and a multitude of supporting factors (Fragkos *et al.*, 2015; Leman and Noguchi, 2013). Due to CMG-catalyzed DNA unwinding ahead of polymerase activity, replisomes form a 'fork-like' structure, commonly known as the replication fork. These processive structures must traverse the DNA fiber, faithfully synthesizing daughter DNA

molecules until fork convergence and termination (Bailey et al., 2015; Petryk *et al.*, 2016).

The elongating replication fork is a vulnerable structure. When intact, the fork contains small tracts of single-stranded DNA (ssDNA), which are quickly converted to double-stranded DNA molecules (dsDNA) upon daughter strand synthesis (Leman and Noguchi, 2013) (Figure 1, top). However, when replication forks encounter physical impediments on the DNA template, they can become stalled (Saxena and Zou, 2022). Fork stalling, in turn, leads to the generation of large tracts of ssDNA, typically through uncoupling of the replicative helicase and polymerases (Dobbelstein and Sorensen, 2015; Saxena and Zou, 2022; Toledo et al., 2017) (Figure 1, bottom). Stalled forks and subsequent ssDNA formation is collectively known as ‘replication stress’, and these aberrant structures activate what is known as the replication stress response. Briefly, exposed ssDNA directly recruits the heterotrimeric replication protein A (RPA), which in turn recruits the ATR kinase. ATR, upon binding RPA, phosphorylates several substrates, including Chk1 to activate the intra-S phase checkpoint (Dobbelstein and Sorensen, 2015; Toledo *et al.*, 2017) (Figure 1, left). The signaling cascade induced by excess ssDNA leads to the inhibition of new RIS firing, fork protection and re-start, and delay of mitotic entry (Toledo *et al.*, 2017). Such events enable RPA pools to be conserved, replication forks to be repaired and restored, and genome replication to still be completed prior to mitosis despite challenges. This ultimately results in maintenance of genome stability and cell viability.

The RPA-ATR axis has the capacity to buffer replication stress at mild to moderate levels. This is likely due to the estimated 6-10 fold excess RPA present in a cell under unstressed conditions (Toledo *et al.*, 2017). However, at high levels of stress, RPA pools become completely sequestered by exposed ssDNA, leading to the accumulation of non-coated forks (Toledo *et al.*, 2017; Toledo et al., 2013) (Figure 1, right). Under these conditions, an event known as replication catastrophe (RC) occurs. In RC, non-coated forks are processed by endonucleases, such as MUS81 (Matos et al., 2020; Toledo *et al.*, 2017), into highly toxic double-stranded breaks (DSBs). DSB-mediated signaling then activates apoptotic or senescent programs, leading to cell death or growth arrest (Dobbelstein and Sorensen, 2015; Norbury and Zhivotovsky, 2004). Thus, suppressing replication fork stalling events during S-phase is critical for RPA pool conservation and cell viability. The formation of highly stable, aberrant fork structures could lead to RC.

Transcription-replication collisions

Transcription and replication occur simultaneously on the genome. Both processes are driven by large polymerase-containing complexes which simultaneously open duplex DNA and traverse the DNA fiber. The co-existence of these bulky, processive structures suggests that transcription and replication could potentially interfere with each other. Given the genotoxic consequences of replication fork stalling, there has been much interest garnered over the years in determining the consequences of transcription-replication collisions (TRCs) on fork processivity. As a result, a model of TRC outcomes can be constructed from both in-vitro and in-vivo work (Figure 2) (Bruning and Marians,

2020; Hamperl *et al.*, 2017; Kumar *et al.*, 2021; Lee et al., 2020; Prado and Aguilera, 2005).

TRCs can potentially occur in four different contexts (Figure 2). Spatially, TRCs can occur in a co-directional manner, in which the replisome and RNA polymerase II (RNAPII) travel in the same direction (Figure 2, top 2 panels), or head-on, in which they converge (Figure 2, bottom 2 panels). Moreover, from a sequence standpoint, TRCs can either occur over C-rich template strands that encourage re-hybridization of nascent RNA (R-loop forming sequence, or RLFS) (Figure 2, right 2 panels), or non-C-rich strands where re-annealing is absent (Figure 2, left 2 panels) (Aguilera and Garcia-Muse, 2012; Helmrich *et al.*, 2013). Re-annealing of RNA to C-rich DNA forms a three-stranded nucleic acid structure known as an R-loop (Aguilera and Garcia-Muse, 2012). In these structures, the G-rich non-template strand can form intra-strand secondary structures known as G-quadruplexes (G4s) (Figure 2, right 2 panels) (Kumar *et al.*, 2021; Lee *et al.*, 2020). Thus, both the spatial and sequence context of a collision must be considered when assessing outcomes.

Interestingly, in all but one context, collisions do not stably stall replication forks in reconstituted systems in-vitro. When collisions are stimulated co-directionally in in-vitro eukaryotic and bacterial systems, CMG helicase directly removes RNAPII from the DNA template, enabling replisome continuation (Figure 2, upper-left) (Bruning and Marians, 2020; Kumar *et al.*, 2021). This ability to bypass co-directional RNAPII occurs regardless of R-loops, as CMG helicase can directly remove R-loops via unwinding, and

lagging strand synthesis can ‘jump over’ opposite strand G4s, leaving a ssDNA gap that can be filled during mitosis (Figure 2, upper-right) (Bruning and Marians, 2020; Kumar *et al.*, 2021). This finding has been re-capitulated in-vivo, where induced co-directional collisions were found to remove R-loops and occur without genotoxic consequence. Similarly, head-on transcription-replication collisions (HO TRCs) over non-C-rich RNAPII template strands failed to incur stable replication fork stalling in-vitro, or DNA damage in-vivo (Figure 2, bottom-left) (Hamperl *et al.*, 2017; Kumar *et al.*, 2021). Thus, most potential collisions that could occur on the genome appear to not cause replication fork stalling and are well tolerated by the cell.

Alternatively, when HO TRCs are stimulated in-vitro over C-rich sequences in the template strand of RNA polymerase II (RNAPII), the replication fork becomes stably blocked (Figure 2, bottom-right) (Bruning and Marians, 2020; Kumar *et al.*, 2021). In this sequence context, nascent RNA from head-on RNAPII forms an R-loop, thus stabilizing G4s on the replisome’s leading strand. While CMG helicase can bypass the G4 on the leading strand, replicative polymerase cannot, leading to uncoupling and ssDNA formation (Figure 3). Alternatively, CMG helicase can be blocked by highly stable G4s, leading to fork reversal, MUS81-dependent cleavage, and ssDNA generation in this context (Kumar *et al.*, 2021; Matos *et al.*, 2020). It is likely that G4 stabilization contributes to fork block, as introduction of the G4 unwinding helicase PIF1 rescues fork stalling (Kumar *et al.*, 2021). In agreement with these findings, in-vivo induction of HO TRCs over RLFS generates R-loops at the collision site and concomitant DNA damage (Hamperl *et al.*, 2017; Nojima et al., 2018). The potent damage induction observed in-

vivo suggests that HO TRCs over RLFS generate stably stalled forks that eventually are converted into DNA breaks. Collectively, these fascinating studies demonstrate that HO TRCs over C-rich R-loop forming sequences (RLFS) (genotoxic TRCs) potently block replication fork progression in a conserved manner.

Transcription-replication coordination

The 3 billion base-pair genome plays host to a myriad of metabolic processes, including DNA replication and RNA transcription. It is currently estimated that there are ~20,000 protein-coding genes on the genome, as well as ~50,000 non-coding transcription units termed lincRNAs (Hangauer et al., 2013; Pertea et al., 2018). Furthermore, ~85% of the genome is transcribed at detectable levels (Hangauer *et al.*, 2013; Jacquier, 2009). Alternatively, it is estimated that ~20,000-100,000 replication forks are active in any given S-phase, initiating from around 10,000-50,000 initiation sites (Fragkos *et al.*, 2015; Ganier *et al.*, 2019; Kumagai and Dunphy, 2020). Replication initiation and active transcription cluster in genomic space (Chen *et al.*, 2019; Dellino *et al.*, 2013; Langley *et al.*, 2016). The ubiquitous and proximal nature of both processes suggests that the avoidance of genotoxic TRCs stems from directional or temporal coordination.

One-way genotoxic TRCs could be avoided in crowded genomic space is enforced co-directionality between all transcription units and replication forks. The development of OK-seq, which maps replication fork directionality genomewide, has helped decipher this relationship (Petryk *et al.*, 2016). OK-seq datasets from HeLa and lymphoid blastoma cells reveal a conserved replication landscape, in which replication initiates

upstream of active genes in ~30kb regions dubbed initiation zones (IZs), elongates co-directionally with coding transcription through gene bodies, and terminates downstream of genes in termination zones (TZs) (Petryk *et al.*, 2016). Orthogonal approaches such as optical replication mapping, have unveiled a similar replication landscape via a small molecule approach (Wang *et al.*, 2021). Interestingly, when a strong promoter was ectopically placed into the genome, it directed the formation of an upstream IZ, suggesting that transcription units themselves direct replication activity, likely through the formation of strong NDRs in the promoter region (Chen *et al.*, 2019). These studies to date have argued for a 'passive' model in which genotoxic TRCs are passively mitigated through genome-encoded co-directionality between both processes (Figure 4).

Alternatively, genotoxic TRCs could be avoided through the silencing of head-on transcription units (HO TUs) during genome replication (Figure 5). Several lines of evidence support this 'active' model of coordination. The passive model only accounts for interference from protein-coding gene transcription, which occurs across only ~3% of the genome (Hangauer *et al.*, 2013). It is estimated that 75-90% of the genome is transcribed, and that a majority of this transcription is non-coding, or 'pervasive' in nature (Jacquier, 2009). Indeed, ~50,000 pervasive transcription units (pervasive TUs, also referred to as lincRNAs) have been called from RNA-seq data, demonstrating that non-gene loci are transcribed with moderate frequency (Hangauer *et al.*, 2013). These, pervasive TUs are often transcribed antisense to genes, suggesting they could provide a source of head-on transcription (Hangauer *et al.*, 2013; Jacquier, 2009). Moreover,

pervasive transcription was found to initiate at origin-of-replication complex bound loci in HeLa cells, in an accessible region that overlaps with C-rich DNA and G4 quadruplex forming sequences (Dellino *et al.*, 2013). However, the directionality of this transcription relative to the emerging replication fork remains unclear. Additionally, the passive model does not account for intragenic RIS, which would produce replication forks that converge with gene transcription. It is clear intragenic RIS are abundant on the genome, as they have been mapped across different methodologies (Dellino *et al.*, 2013; Langley *et al.*, 2016; Petryk *et al.*, 2016). Interestingly, most intragenic RIS localize just downstream of the TSS, where abortive TSS transcription would be highly active.

Functional studies suggest that transcriptional silencing is necessary to prevent genotoxic TRCs in human cells. Knockdown of the positive elongation factor Spt6 increased the expression of promoter-upstream transcripts (PROMPTs) in asynchronous HeLa cells, leading to the build up of NET-seq signal at select upstream origins, PROMPT-associated R-loops and DNA damage, and replication stress phenotypes (Nojima *et al.*, 2018). Inhibition of the positive elongation factor BRD4 similarly generated evidence of genotoxic TRCs in HeLa and HCT116 cells (Lam et al., 2020). These studies suggest that transcriptional regulation by multiple players is necessary to avoid genotoxic TRCs. However, as these studies were only performed at select loci in asynchronous cells, it remains unknown how widespread head-on transcription over RLFS occurs on the genome, or whether such transcription is temporally regulated. Collectively, the widespread presence of pervasive TUs, existence of intragenic RIS, and evidence of genotoxic TRC suppression by regulators supports

an ‘active’ model of transcription-replication coordination. However, this model has not been systematically tested on a global scale with temporal resolution.

Induction of genotoxic TRCs as a strategy to treat cancer

A hallmark of cancer is elevated replication stress (Gaillard et al., 2015). This phenotype is oncogene-driven, resulting in a myriad of stress-causing phenomena such as depleted dNTP pools, increased nucleotide alterations, and compromised cell cycle checkpoints (Dobbelstein and Sorensen, 2015; Saxena and Zou, 2022). These collectively lead to increased endogenous levels of exposed ssDNA and decreased available RPA (Toledo *et al.*, 2017; Toledo *et al.*, 2013). It has been suggested that tumor-specific replication stress can be exploited therapeutically (Dobbelstein and Sorensen, 2015; Ubhi and Brown, 2019). If tumor cells have less ability to buffer increases in ssDNA relative to healthy cells due to depleted RPA pools, then perturbations that generate novel sources of replication stress, through, for example, inducing genotoxic TRCs, might achieve tumor-selective toxicity (Figure 6). In support of this, BRD4 inhibition across cell lines was shown to selectively generate DNA damage in oncogene-driven models with high stress phenotypes (Lam *et al.*, 2020). Moreover, several drugs that function through generating stress via intra or inter-strand DNA crosslinks, protein-DNA crosslinks, or base modifications, serve as the backbones for approved cancer therapy regimens (Dobbelstein and Sorensen, 2015). However, many chemotherapies induce replication stress-independent toxicities as well, leading to undesirable side-effect profiles and dose limitations (Barabas et al., 2008). Furthermore, new downstream approaches targeting the replication stress-response, such as CHK1

inhibition, have been found to be highly toxic (Dent, 2019). The discovery of novel mechanisms that function to suppress sources of replication stress in tumor cells could potentially lead to the development of next-generation therapeutics with enhanced profiles relative to the current standards of care. Transcriptional regulators that suppress genotoxic TRCs are attractive targets in this regard.

INO80 and MOT1

INO80C is a highly conserved, multi-functional 15-subunit chromatin remodeling complex that participates in transcription, replication, and DNA repair (Poli et al., 2017). INO80 affects DNA processes through its ability to mobilize nucleosomes, perform histone variant exchange, and directly interact with trans factors (Brahma et al., 2017; Lafon et al., 2015; Poli *et al.*, 2017). MOT1 is a highly conserved ATPase that functions as a TATA-binding protein (TBP) antagonist (Auble et al., 1994). INO80 and MOT1 have a unique function amongst transcriptional regulators in that they silence pervasive transcription across model organisms (Xue et al., 2017). For example, in both yeast and mouse embryonic stem cells, INO80 and MOT1 were found to bind at gene TSS. Upon INO80 and MOT1 co-depletion, upstream antisense RNA transcription from bound TSS greatly increased, with little effect on gene expression observed (Xue *et al.*, 2017). While it was originally unclear why INO80 and MOT1 would silence non-productive transcription in a conserved manner, it was later found that INO80 and MOT1 selectively bound and silenced pervasive transcription near RIS (Topal et al., 2020). Indeed, co-depletion in yeast under induced replication stress generated DNA breaks at RIS, suggesting that INO80 and MOT1 function to prevent genotoxic TRCs through

silencing pervasive transcription in this model organism (Topal *et al.*, 2020). However, it is unclear if INO80 and MOT1 regulate pervasive transcription or prevent genotoxic TRCs in human cells (Figure 7).

INO80C has recently emerged as an oncogenic protein complex in various cancers (Lee et al., 2017; Zhang et al., 2017; Zhou et al., 2016). Pan-tumor analysis of sequencing and expression data from the cancer genome atlas (TCGA) revealed that INO80 is genetically amplified across most cancer subtypes (Lee *et al.*, 2017; Zhang *et al.*, 2017). INO80 depletion in melanoma, prostate, breast, colorectal, and lung cancer cell lines inhibits growth, suggesting that INO80 has pan-cancer oncogenic activity (Lee *et al.*, 2017; Prendergast et al., 2020; Zhang *et al.*, 2017; Zhou *et al.*, 2016). However, it is unclear how INO80 mechanistically facilitates tumor expansion. Work in melanoma and NSCLC models suggested that INO80 drives oncogenic enhancer activation, although such studies were descriptive in nature (Zhang *et al.*, 2017; Zhou *et al.*, 2016). Alternatively, work in colorectal, prostate, and breast cancer models demonstrated that INO80 suppresses replication stress and replication-dependent DNA damage through an unclear mechanism, (Lee *et al.*, 2017; Prendergast *et al.*, 2020). Interestingly, INO80 depletion was found to be synthetic lethal with replication stress in yeast, suggesting that INO80 could be interacting with a tumor-specific stress phenotype (Papamichos-Chronakis and Peterson, 2008). Could INO80 be functioning to prevent genotoxic TRCs in these cancer types? Interestingly, work in PC3 cells found that INO80 increases chromatin occupancy in S-phase, supporting the idea that INO80 could be an S-phase specific transcriptional regulator (Vassileva et al., 2014). Furthermore, INO80 depletion

increased R-loop occupancy on chromatin in S-phase PC3 cells, a phenotype reflecting upregulated genotoxic TRCs (Prendergast *et al.*, 2020). However, INO80 regulatory activity was not assessed at high resolution in these studies. Thus, it is unclear how INO80 mechanistically prevents DNA damage during S-phase in tumor cells.

Non-Small Cell Lung cancer

Lung cancer is currently the leading cause of cancer-related deaths for both men and women in the United States (Herbst et al., 2018). Non-Small Cell Lung cancer (NSCLC) is the most prevalent histologic lung cancer subtype, accounting for ~85% of total cases (Herbst *et al.*, 2018). First line treatment for advanced NSCLC is typically cisplatin-based chemotherapy, which produces a ~20% response rate and ~8-month median survival across different regimens (Fennell et al., 2016). In a small subset of NSCLC tumors harboring EGFR or ALK mutations, targeted kinase inhibitors can significantly prolong survival, demonstrating improved efficacy over cisplatin-based regimens (Alanazi et al., 2020). However, these treatments are rarely curative, with most patients progressing on therapy. Thus, there is a clear need to develop novel therapies for NSCLC.

NSCLC is unique amongst cancer types in that it displays remarkable levels of genome instability. In support of this, pan-cancer analysis found that NSCLC exhibits a rapid mutation rate of ~50 mutations/Mb of DNA, the highest amongst cancers assessed (Kandoth et al., 2013). A priori, it can be assumed that this instability is driven afferently by replication stress, and efferently by compromised DNA damage response and repair

mechanisms. Indeed, NSCLC cell line panels have shown uniquely high levels of ssDNA, a strong biomarker for replication stress (Boucher et al., 2019; Zhao et al., 2009). NSCLC genetically exhibits high rates of KRAS gain-of-function mutations, which have been shown to induce replication stress in controlled systems (Araujo et al., 2021; Kotsantis et al., 2016). Moreover, NSCLC exhibits a high rate of inactivating mutations in the SWI/SNF subunit SMARCA4, which likewise drive heterochromatin-induced replication stress in controlled systems (Kurashima et al., 2020; Medina et al., 2008). Thus, NSCLC profiles as a strong candidate tumor subtype for therapies that exogenously introduce replication stress, therefore exploiting the depleted RPA pools in these tumor cells. Interestingly, INO80 has been found to be critical for NSCLC growth across cell lines, although INO80's oncogenic mechanism remains unclear (Zhang *et al.*, 2017). It is striking that INO80 depletion does not affect the growth of normal lung epithelial cells, suggesting that INO80 is interacting with a tumor-specific phenotype (Zhang *et al.*, 2017). Whether INO80 functions to prevent genotoxic TRCs in NSCLC remains unexplored.

Approach

In this thesis work, I set out to answer two main questions: 1. Is transcription actively regulated to avoid genotoxic TRCs in tumor cells? and 2. Do the transcriptional regulators INO80 and MOT1 prevent genotoxic TRCs in NSCLC? To answer the first question, I leveraged a multitude of publicly available datasets, including CAGE-seq, SNS-seq, and phased GRO-seq in the MCF-7 breast cancer cell line to investigate if 1. HO TUs exist on the genome, and 2. HO TUs are silenced during S-phase at potential

genotoxic TRC sites. To broach the second question, I performed both high resolution genomic assays, including ChIP-seq, nascent RNA-seq, EdU-seq, and bulk DNA damage assessments to interrogate if 1. INO80 and MOT1 bound at HO TUs in NSCLC, 2. INO80 and MOT1 silenced HO TU transcription, and 3. INO80 and MOT1 prevented TRC-induced DNA damage. These dual studies, which are both descriptive and functional in nature, enabled me to ultimately ascertain whether genotoxic TRCs are prevented on the cancer genome through transcriptional silencing mechanisms.

Significance

Despite large advances in the last twenty years, there remains a high unmet need for improved cancer therapeutics. The discovery of novel cancer-specific vulnerabilities due to tumor pathology is key to efforts in drug development. This work seeks to illuminate a potential weakness in tumor cells derived from abnormal DNA metabolism. Moreover, the interplay between transcription and replication has been understudied, potentially due to a lack of crosstalk between the transcription and replication fields. This work seeks to bridge this intellectual gap and potentially draw attention to the idea that transcriptional regulatory mechanisms can be critical to cellular physiology independent of gene expression.

Summary

In this work, I will first provide evidence in support of an ‘active’ transcription-replication coordination model through an integrative bioinformatic analysis performed from data generated in MCF-7 breast cancer cells. I will then pivot to presenting data supporting a

model by which the transcriptional regulators INO80 and MOT1 function to suppress genotoxic TRCs in NSCLC, and thus facilitate tumor cell growth. My work will culminate in a descriptive study mapping the genome-wide DNA damage effects induced by cisplatin treatment in NSCLC cells. This work ultimately illuminates a novel paradigm in transcription-replication coordination on the human genome and provides rationale for the therapeutic targeting of such coordination in cancer.

Figures

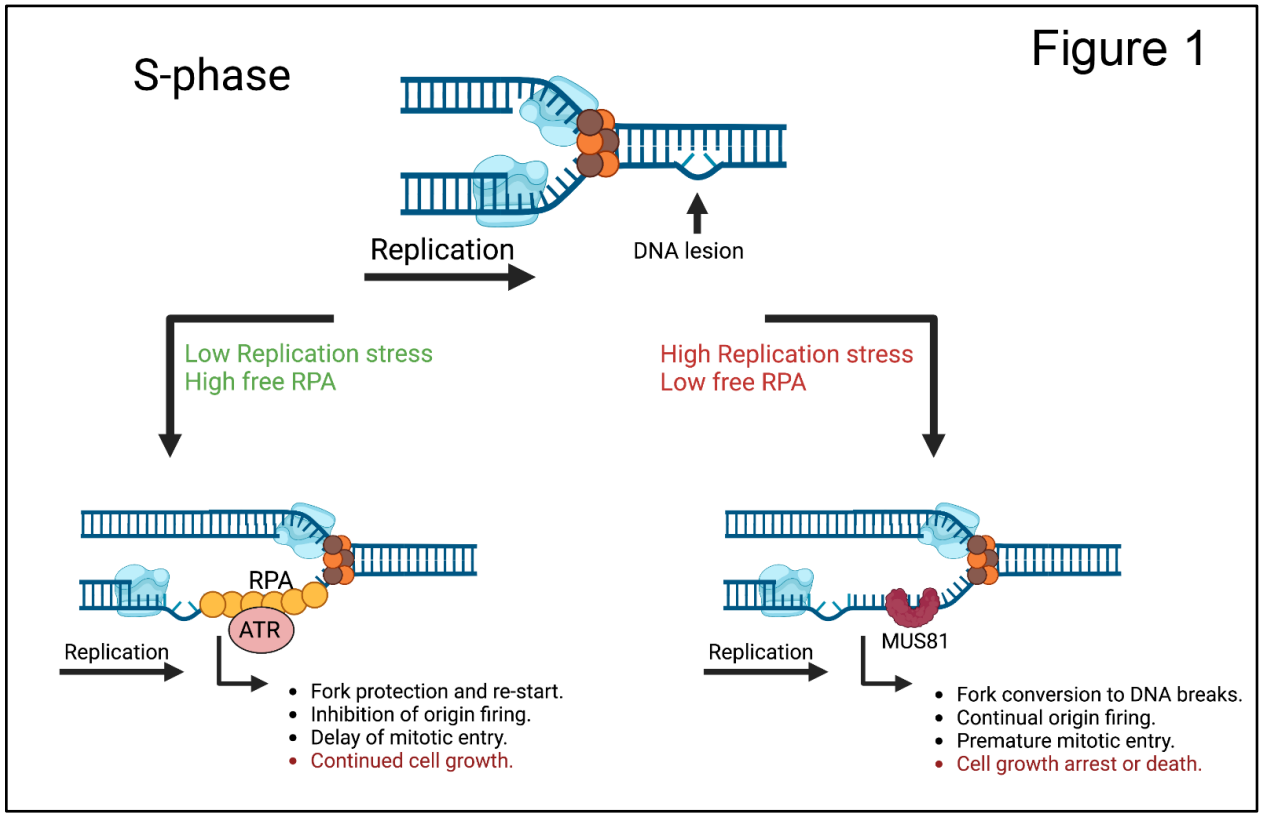


Figure 1
S-phase
Replication
DNA lesion
Low Replication stress
High free RPA
High Replication stress
Low free RPA
RPA
ATR
Replication
Fork protection and re-start.
Inhibition of origin firing.
Delay of mitotic entry.
Continued cell growth.
MUS81
Replication
Fork conversion to DNA breaks.
Continual origin firing.
Premature mitotic entry.
Cell growth arrest or death.

Figure 2

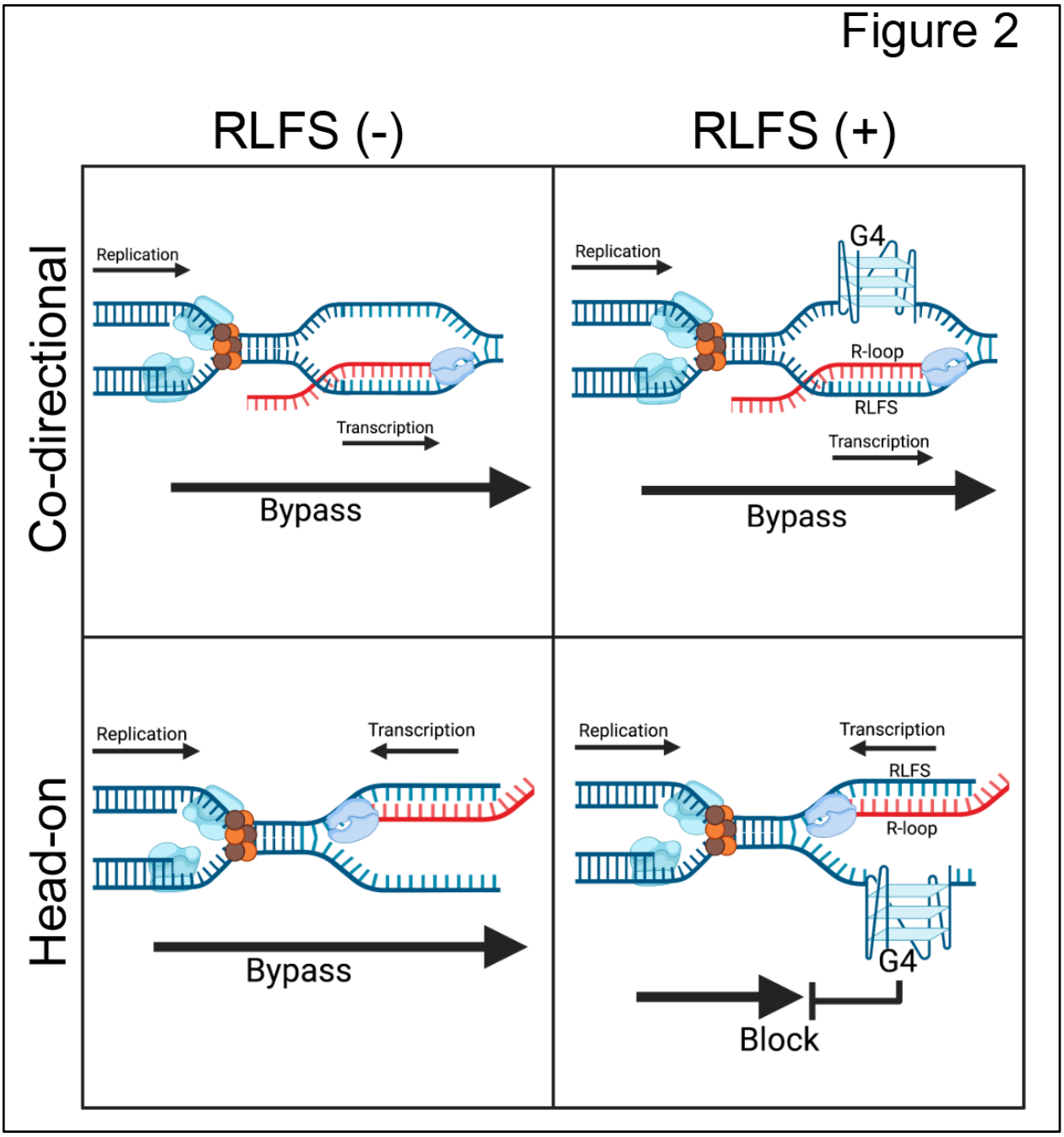
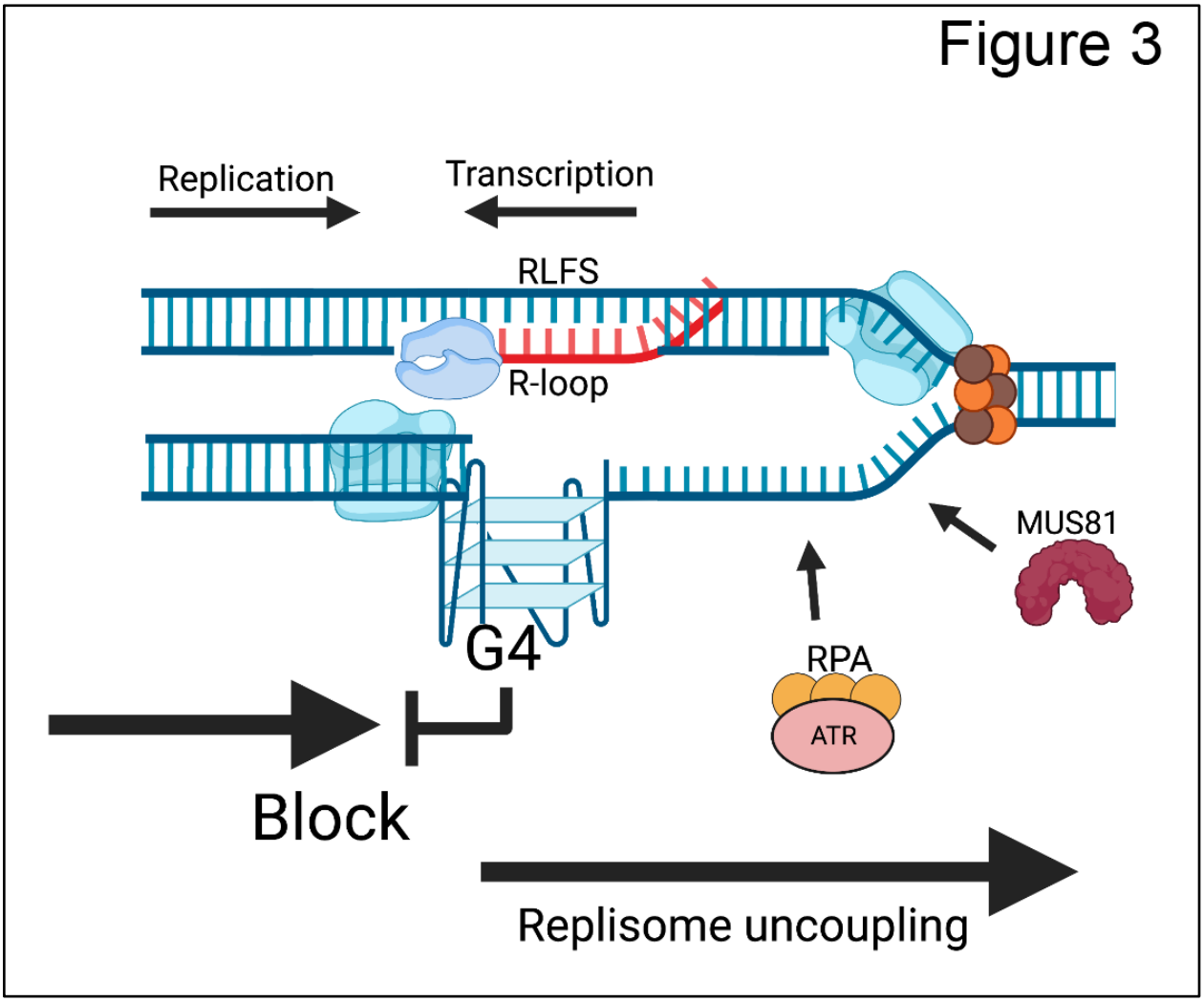

RLFS (-)
RLFS (+)
Co-directional
Head-on
Replication
Transcription
Bypass
G4
R-loop
RLFS
Block

Figure 3


Replication
Transcription
RLFS
R-loop
MUS81
G4
RPA
ATR
Block
Replisome uncoupling

S-phase

Figure 4

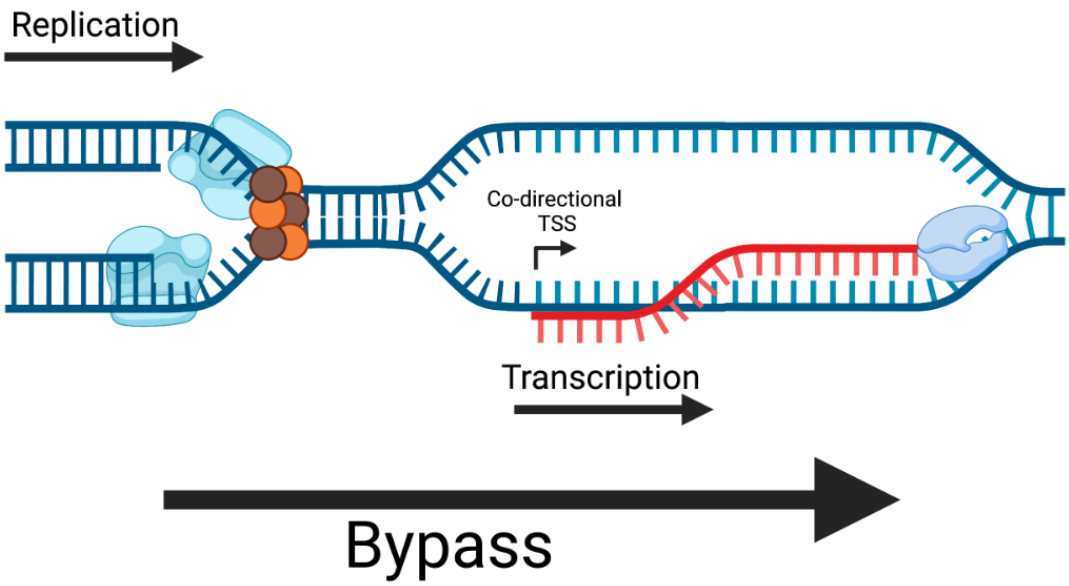


Figure 5

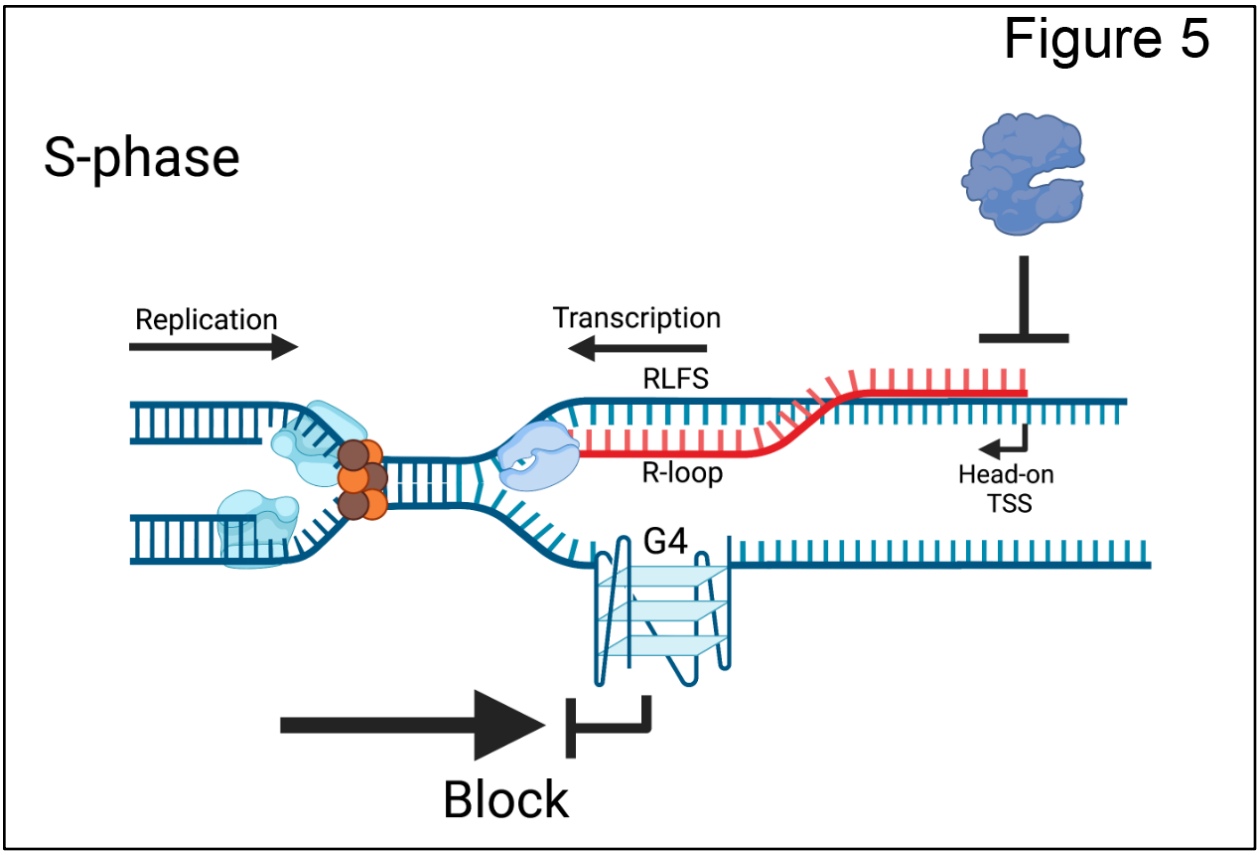

S-phase
Replication
Transcription
RLFS
R-loop
Head-on
TSS
G4
Block

Figure 6

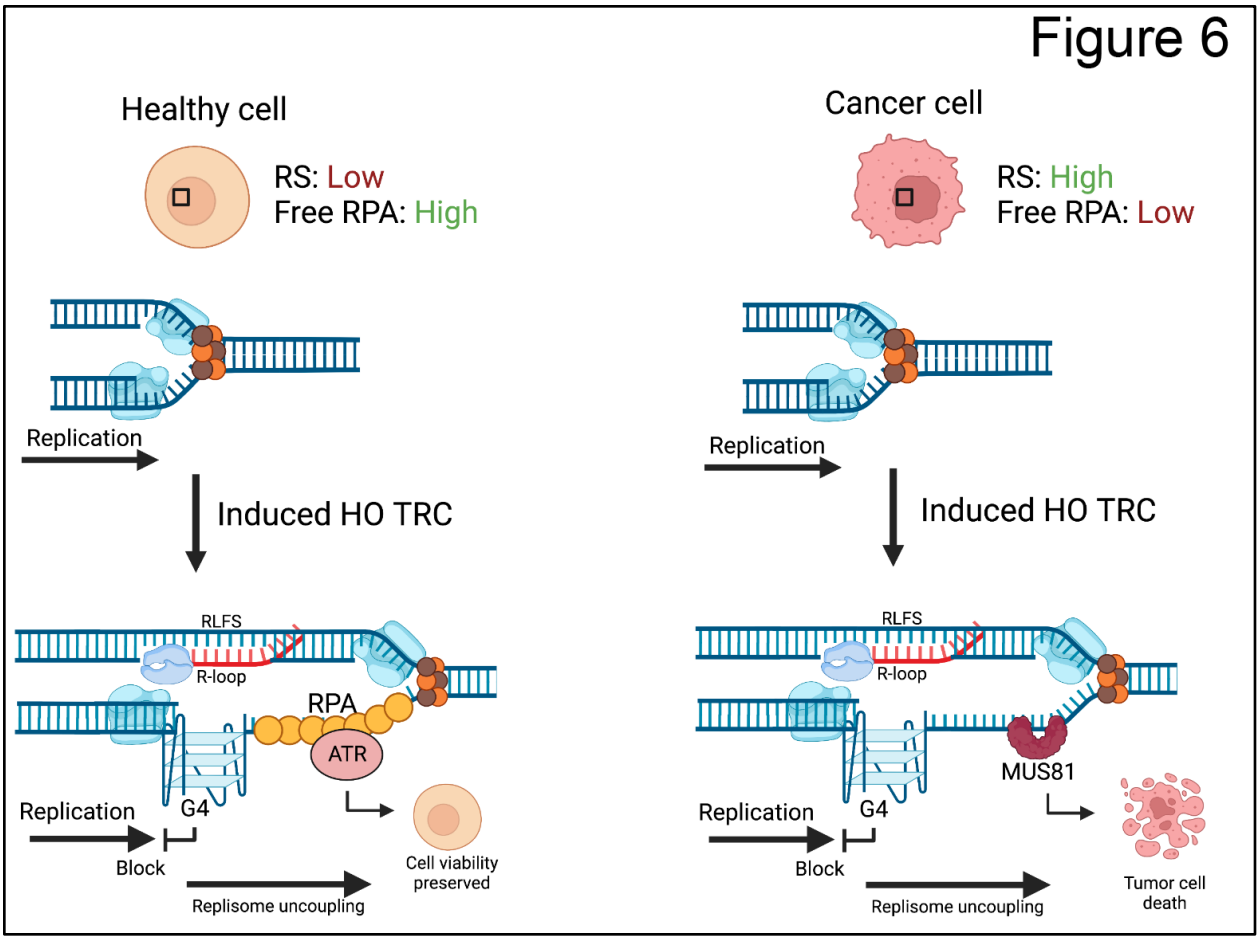

Healthy cell
RS: Low
Free RPA: High
Replication
Induced HO TRC
RLFS
R-loop
RPA
ATR
Replication
G4
Block
Replisome uncoupling
Cell viability preserved
Cancer cell
RS: High
Free RPA: Low
Replication
Induced HO TRC
RLFS
R-loop
MUS81
Replication
G4
Block
Replisome uncoupling
Tumor cell death

Figure 7

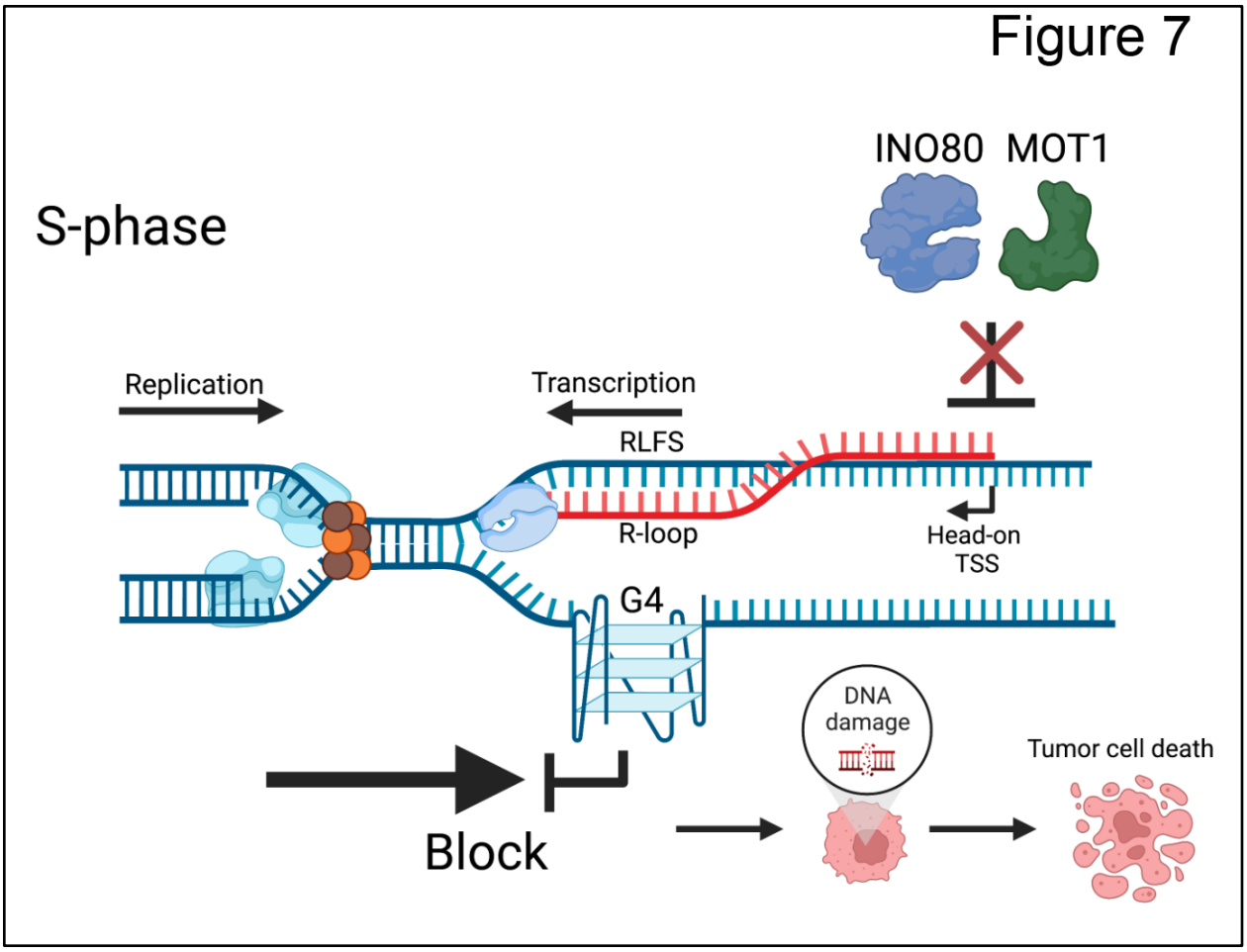

S-phase
INO80
MOT1
Replication
Transcription
RLFS
R-loop
Head-on
TSS
G4
DNA
damage
Tumor cell death
Block

Figure Legends

Figure 1: Consequences of replication fork stalling. Graphic representation of cellular response to replication fork stalling in conditions of low replication stress (left) and high replication stress (right).

Figure 2: Outcomes of transcription-replication collisions in different contexts. Graphic representation of co-directional collisions at sites with no R-loop forming sequences (upper-left), co-directional collisions at sites with R-loop forming sequences (upper-right), head-on collisions at sites with no R-loop forming sequences (lower-left), and head-on collisions at sites with R-loop forming sequences (lower-right). In the first 3 contexts, the replication fork does not stall and is able to continue synthesis. In the final context, the replisome becomes stably blocked due to the formation of G4 secondary structures on the leading strand.

Figure 3: Consequences of replication fork stalling induced by genotoxic transcription-replication collisions. Graphic representation of replication fork stalling induced by a head-on transcription-replication collision over an R-loop forming sequence. In this scenario, leading strand synthesis is blocked by a G4 quadruplex, which the CMG helicase is able to bypass. This leads to replisome uncoupling and single-stranded DNA generation. Lagging strand synthesis is capable of bypassing the transcriptional complex to continue DNA synthesis.

Figure 4: Passive model of transcription-replication coordination. Graphic representation of transcription-replication coordination in a 'passive' system in which all transcription-replication collisions are co-directional in nature. This directional relationship would enable replisome bypass and continual synthesis without fork stalling.

Figure 5: Active model of transcription-replication coordination. Graphic representation of transcription-replication coordination in an 'active' system in which head-on transcription occurs on the genome but is silenced during S-phase by transcriptional regulators. Temporal suppression of head-on transcripts would enable replisome passage during S-phase.

Figure 6: Rationale for therapeutically targeting genotoxic collisions in cancer. Graphic representation of the consequences of inducing genotoxic transcription-replication collisions in healthy cells (left) and tumor cells (right). Due to increased endogenous replication stress and depleted RPA pools in tumor cells, induced collisions would generate unstable forks that could be converted into DNA breaks by endonucleases such as MUS81, leading to tumor cell death or growth arrest.

Figure 7: Model of INO80 and MOT1's potential protective function in cancer. Graphic representation of a molecular model in which INO80 and MOT1 function to silence head-on transcription, thus suppressing genotoxic collisions and preserving

tumor cell viability. Inhibition of the INO80/MOT1 axis would lead to collision induction, DNA damage, and downstream cytotoxic effects in tumor cells.

References

Aguilera, A., and Garcia-Muse, T. (2012). R loops: from transcription byproducts to threats to genome stability. Mol Cell *46*, 115-124. 10.1016/j.molcel.2012.04.009.

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. 10.1038/s41467-020-18527-0.

Alanazi, A., Yunusa, I., Elenizi, K., and Alzarea, A.I. (2020). Efficacy and safety of tyrosine kinase inhibitors in advanced non-small-cell lung cancer harboring epidermal growth factor receptor mutation: a network meta-analysis. Lung Cancer Manag *10*, LMT43. 10.2217/lmt-2020-0011.

Araujo, L.H., Souza, B.M., Leite, L.R., Parma, S.A.F., Lopes, N.P., Malta, F.S.V., and Freire, M.C.M. (2021). Molecular profile of KRAS G12C-mutant colorectal and non-small-cell lung cancer. BMC Cancer *21*, 193. 10.1186/s12885-021-07884-8.

Auble, D.T., Hansen, K.E., Mueller, C.G., Lane, W.S., Thorner, J., and Hahn, S. (1994). Mot1, a global repressor of RNA polymerase II transcription, inhibits TBP binding to DNA by an ATP-dependent mechanism. Genes Dev *8*, 1920-1934. 10.1101/gad.8.16.1920.

Bailey, R., Priego Moreno, S., and Gambus, A. (2015). Termination of DNA replication forks: "Breaking up is hard to do". Nucleus *6*, 187-196. 10.1080/19491034.2015.1035843.

Barabas, K., Milner, R., Lurie, D., and Adin, C. (2008). Cisplatin: a review of toxicities and therapeutic applications. Vet Comp Oncol *6*, 1-18. 10.1111/j.1476-5829.2007.00142.x.

Boucher, D., Ashton, N., Suraweera, A., Burgess, J., Bolderson, E., Barr, M., Gray, S., Gately, K., Adams, M., Croft, L., et al. (2019). Human single-stranded DNA protein 1 (hSSB1): a prognostic factor and target for non-small cell lung cancer (NSCLC) treatment. Lung Cancer *127*, S17-S17. Doi 10.1016/S0169-5002(19)30084-4.

Brahma, S., Udugama, M.I., Kim, J., Hada, A., Bhardwaj, S.K., Hailu, S.G., Lee, T.H., and Bartholomew, B. (2017). INO80 exchanges H2A.Z for H2A by translocating on DNA proximal to histone dimers. Nat Commun *8*, 15616. 10.1038/ncomms15616.

Bruning, J.G., and Marians, K.J. (2020). Replisome bypass of transcription complexes and R-loops. Nucleic Acids Res *48*, 10353-10367. 10.1093/nar/gkaa741.

Chen, Y.H., Keegan, S., Kahli, M., Tonzi, P., Fenyo, 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. 10.1038/s41594-018-0171-0.

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. 10.1101/gr.142331.112.

Dent, P. (2019). Investigational CHK1 inhibitors in early phase clinical trials for the treatment of cancer. Expert Opin Investig Drugs *28*, 1095-1100. 10.1080/13543784.2019.1694661.

Dobbelstein, M., and Sorensen, C.S. (2015). Exploiting replicative stress to treat cancer. Nat Rev Drug Discov *14*, 405-423. 10.1038/nrd4553.

Fennell, D.A., Summers, Y., Cadranel, J., Benepal, T., Christoph, D.C., Lal, R., Das, M., Maxwell, F., Visseren-Grul, C., and Ferry, D. (2016). Cisplatin in the modern era: The backbone of first-line chemotherapy for non-small cell lung cancer. Cancer Treat Rev *44*, 42-50. 10.1016/j.ctrv.2016.01.003.

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. 10.1101/gr.183848.114.

Fragkos, M., Ganier, O., Coulombe, P., and Mechali, M. (2015). DNA replication origin activation in space and time. Nat Rev Mol Cell Biol *16*, 360-374. 10.1038/nrm4002.

Gaillard, H., Garcia-Muse, T., and Aguilera, A. (2015). Replication stress and cancer. Nat Rev Cancer *15*, 276-289. 10.1038/nrc3916.

Ganier, O., Prorok, P., Akerman, I., and Mechali, M. (2019). Metazoan DNA replication origins. Curr Opin Cell Biol *58*, 134-141. 10.1016/j.ceb.2019.03.003.

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 e719. 10.1016/j.cell.2017.07.043.

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. 10.1371/journal.pgen.1003569.

Helmrich, A., Ballarino, M., Nudler, E., and Tora, L. (2013). Transcription-replication encounters, consequences and genomic instability. Nat Struct Mol Biol *20*, 412-418. 10.1038/nsmb.2543.

Herbst, R.S., Morgensztern, D., and Boshoff, C. (2018). The biology and management of non-small cell lung cancer. Nature *553*, 446-454. 10.1038/nature25183.

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. 10.1074/jbc.M113.492504.

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

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333-339. 10.1038/nature12634.

Kotsantis, P., Silva, L.M., Irmscher, S., Jones, R.M., Folkes, L., Gromak, N., and Petermann, E. (2016). Increased global transcription activity as a mechanism of replication stress in cancer. Nat Commun *7*, 13087. 10.1038/ncomms13087.

Kumagai, A., and Dunphy, W.G. (2020). Binding of the Treslin-MTBP Complex to Specific Regions of the Human Genome Promotes the Initiation of DNA Replication. Cell Rep *32*, 108178. 10.1016/j.celrep.2020.108178.

Kumar, C., Batra, S., Griffith, J.D., and Remus, D. (2021). The interplay of RNA:DNA hybrid structure and G-quadruplexes determines the outcome of R-loop-replisome collisions. Elife *10*. 10.7554/eLife.72286.

Kurashima, K., Kashiwagi, H., Shimomura, I., Suzuki, A., Takeshita, F., Mazevet, M., Harata, M., Yamashita, T., Yamamoto, Y., Kohno, T., and Shiotani, B. (2020). SMARCA4 deficiency-associated heterochromatin induces intrinsic DNA replication stress and susceptibility to ATR inhibition in lung adenocarcinoma. NAR Cancer *2*, zcaa005. 10.1093/narcan/zcaa005.

Lafon, A., Taranum, S., Pietrocola, F., Dingli, F., Loew, D., Brahma, S., Bartholomew, B., and Papamichos-Chronakis, M. (2015). INO80 Chromatin Remodeler Facilitates Release of RNA Polymerase II from Chromatin for Ubiquitin-Mediated Proteasomal Degradation. Mol Cell *60*, 784-796. 10.1016/j.molcel.2015.10.028.

Lam, F.C., Kong, Y.W., Huang, Q., Vu Han, T.L., Maffa, A.D., Kasper, E.M., and Yaffe, M.B. (2020). BRD4 prevents the accumulation of R-loops and protects against transcription-replication collision events and DNA damage. Nat Commun *11*, 4083. 10.1038/s41467-020-17503-y.

Langley, A.R., Graf, 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. 10.1093/nar/gkw760.

Lee, C.Y., McNerney, C., Ma, K., Zhao, W., Wang, A., and Myong, S. (2020). R-loop induced G-quadruplex in non-template promotes transcription by successive R-loop formation. Nat Commun *11*, 3392. 10.1038/s41467-020-17176-7.

Lee, S.A., Lee, H.S., Hur, S.K., Kang, S.W., Oh, G.T., Lee, D., and Kwon, J. (2017). INO80 haploinsufficiency inhibits colon cancer tumorigenesis via replication stress-induced apoptosis. Oncotarget *8*, 115041-115053. 10.18632/oncotarget.22984.

Leman, A.R., and Noguchi, E. (2013). The replication fork: understanding the eukaryotic replication machinery and the challenges to genome duplication. Genes (Basel) *4*, 1-32. 10.3390/genes4010001.

Matos, D.A., Zhang, J.M., Ouyang, J., Nguyen, H.D., Genois, M.M., and Zou, L. (2020). ATR Protects the Genome against R Loops through a MUS81-Triggered Feedback Loop. Mol Cell *77*, 514-527 e514. 10.1016/j.molcel.2019.10.010.

Medina, P.P., Romero, O.A., Kohno, T., Montuenga, L.A., Pio, R., Yokota, J., and Sanchez-Cespedes, M. (2008). Frequent BRG1/SMARCA4-inactivating mutations in human lung cancer cell lines. Hum Mutat *29*, 617-622. 10.1002/humu.20730.

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 e977. 10.1016/j.molcel.2018.10.011.

Norbury, C.J., and Zhivotovsky, B. (2004). DNA damage-induced apoptosis. Oncogene *23*, 2797-2808. 10.1038/sj.onc.1207532.

Papamichos-Chronakis, M., and Peterson, C.L. (2008). The Ino80 chromatin-remodeling enzyme regulates replisome function and stability. Nat Struct Mol Biol *15*, 338-345. 10.1038/nsmb.1413.

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F.P., Chang, Y.C., Madugundu, A.K., Pandey, A., and Salzberg, S.L. (2018). CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biol *19*, 208. 10.1186/s13059-018-1590-2.

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. 10.1038/ncomms10208.

Poli, J., Gasser, S.M., and Papamichos-Chronakis, M. (2017). The INO80 remodeller in transcription, replication and repair. Philos Trans R Soc Lond B Biol Sci *372*. 10.1098/rstb.2016.0290.

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

Prendergast, L., McClurg, U.L., Hristova, R., Berlinguer-Palmini, R., Greener, S., Veitch, K., Hernandez, I., Pasero, P., Rico, D., Higgins, J.M.G., et al. (2020). Resolution of R-loops by INO80 promotes DNA replication and maintains cancer cell proliferation and viability. Nat Commun *11*, 4534. 10.1038/s41467-020-18306-x.

Saxena, S., and Zou, L. (2022). Hallmarks of DNA replication stress. Mol Cell *82*, 2298-2314. 10.1016/j.molcel.2022.05.004.

Toledo, L., Neelsen, K.J., and Lukas, J. (2017). Replication Catastrophe: When a Checkpoint Fails because of Exhaustion. Mol Cell *66*, 735-749. 10.1016/j.molcel.2017.05.001.

Toledo, L.I., Altmeyer, M., Rask, M.B., Lukas, C., Larsen, D.H., Povlsen, L.K., Bekker-Jensen, S., Mailand, N., Bartek, J., and Lukas, J. (2013). ATR prohibits replication catastrophe by preventing global exhaustion of RPA. Cell *155*, 1088-1103. 10.1016/j.cell.2013.10.043.

Topal, S., Van, C., Xue, Y., Carey, M.F., and Peterson, C.L. (2020). INO80C Remodeler Maintains Genomic Stability by Preventing Promiscuous Transcription at Replication Origins. Cell Rep *32*, 108106. 10.1016/j.celrep.2020.108106.

Ubhi, T., and Brown, G.W. (2019). Exploiting DNA Replication Stress for Cancer Treatment. Cancer Res *79*, 1730-1739. 10.1158/0008-5472.CAN-18-3631.

Vassileva, I., Yanakieva, I., Peycheva, M., Gospodinov, A., and Anachkova, B. (2014). The mammalian INO80 chromatin remodeling complex is required for replication stress recovery. Nucleic Acids Res *42*, 9074-9086. 10.1093/nar/gku605.

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 e2976. 10.1016/j.molcel.2021.05.024.

Xue, Y., Pradhan, S.K., Sun, F., Chronis, C., Tran, N., Su, T., Van, C., Vashisht, A., Wohlschlegel, J., Peterson, C.L., et al. (2017). Mot1, Ino80C, and NC2 Function Coordinately to Regulate Pervasive Transcription in Yeast and Mammals. Mol Cell *67*, 594-607 e594. 10.1016/j.molcel.2017.06.029.

Zhang, S., Zhou, B., Wang, L., Li, P., Bennett, B.D., Snyder, R., Garantziotis, S., Fargo, D.C., Cox, A.D., Chen, L., and Hu, G. (2017). INO80 is required for oncogenic transcription and tumor growth in non-small cell lung cancer. Oncogene *36*, 1430-1439. 10.1038/onc.2016.311.

Zhao, Z., Xu, L., Shi, X., Tan, W., Fang, X., and Shangguan, D. (2009). Recognition of subtype non-small cell lung cancer by DNA aptamers selected from living cells. Analyst *134*, 1808-1814. 10.1039/b904476k.

Zhou, B., Wang, L., Zhang, S., Bennett, B.D., He, F., Zhang, Y., Xiong, C., Han, L., Diao, L., Li, P., et al. (2016). INO80 governs superenhancer-mediated oncogenic transcription and tumor growth in melanoma. Genes Dev *30*, 1440-1453. 10.1101/gad.277178.115.

Chapter 2: Temporal regulation of head-on transcription at replication initiation sites.

Temporal regulation of head-on transcription at replication initiation sites

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

[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

Abstract

Head-on collisions between the DNA replication machinery and RNA polymerase are potent genotoxic events leading to replication fork stalling, R-loop formation, and DNA breaks. Current models suggest that head-on collisions are avoided through replication initiation site (RIS) placement upstream of active genes, thus ensuring co-orientation of replication fork movement and genic transcription. However, this model does not account for pervasive transcription units, or intragenic replication initiation events. Through mining phased GRO-seq data, and developing a rigorous informatic strategy to identify RIS, we demonstrate that head-on transcription occurs frequently in a breast cancer cell line, and that this transcription is significantly downregulated during S-phase, particularly in regions susceptible to R-loop formation. Collectively, our analysis suggests the existence of a temporally tuned transcriptional regulation mechanism that functions to maintain genome stability.

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 studies across model organisms have shown that head-on collisions between the DNA and RNA polymerases are potent genotoxic events, capable of generating stalled replication forks, R-loops, and double-stranded DNA breaks (Hamperl et al., 2017; Lang et al., 2017; Liu and Alberts, 1995; Mirkin and Mirkin, 2005; Prado and Aguilera, 2005; Zardoni et al., 2021). In contrast, co-directional collisions are tolerated and result in little effect on replisome progression (Hamperl *et al.*, 2017; Liu and Alberts, 1995; Mirkin and Mirkin, 2005; Prado and Aguilera, 2005). Such findings highlight the need for cells to preserve genome stability by employing mechanisms to avoid head-on collisions.

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 (Chen et al., 2019; Petryk et al., 2016). 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 (Wang et al., 2021). The organization suggested by these studies would in theory ensure that leading strand synthesis primarily occurs in a co-directional manner with genic transcription. However, there are several limitations to the model. First, although about 2% of the genome is occupied by protein-coding genes, 75-90% of the genome is transcribed (Consortium, 2012; Hangauer et al., 2013). Non-coding, or pervasive transcription, can occur both outside gene bodies, such as in the form of promoter upstream transcripts (PROMPTs) (Berretta and Morillon, 2009; Preker et al.,

2008), or inside genes, such as intragenic cryptic transcripts (McCauley and Dang, 2022; Smolle and Workman, 2013). Often, these units are transcribed antisense to gene transcription, suggesting they could be a source of head-on collisions (Berretta and Morillon, 2009; Xie et al., 2011). Second, several lines of evidence suggest transcription occurs adjacent to replication initiation sites (RIS) (Candelli et al., 2018; Dellino et al., 2013; Miotto et al., 2016), and indeed might be a functional feature of RIS, acting to recruit the replication machinery (Dellino *et al.*, 2013; Hoshina et al., 2013). 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 transcriptional collisions. Third, several assays mapping RIS have found enrichment of peaks within gene bodies. Ini-seq, which maps RIS via digoxigenin-dUTP incorporation, purification, and sequencing, identified 8,048 peaks within genes in EJ3 cells (Langley et al., 2016). ChIP-seq of ORC1 identified 4,272 peaks within genes in HeLa cells (Dellino *et al.*, 2013). SNS-seq, which maps RIS via isolation of lambda exonuclease-resistant RNA-primed DNA fragments found that peaks primarily localized downstream of the TSS of active genes in MCF-7 cells (Martin et al., 2011). OK-seq in HeLa and GM6990 cells showed that ~20% of IZs localized within active genes (Petryk *et al.*, 2016). Therefore, it appears that a subset of RIS initiate within gene bodies across cell types. Within these genes, it is possible that gene transcription could generate head-on collisions.

Collectively, it appears that both pervasive and genic transcription could be potential sources of head-on collisions. However, in the case of pervasive transcription units, it is

unclear how transcription directionally occurs relative to the moving replication fork. In the case of gene units, it is unclear whether RIS occur within genes that are actively transcribed. If pervasive and genic transcription occurred at a high frequency in the head-on orientation, a key question is how do cells avoid head-on collisions at these locations? 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 3 kilobases of stringently identified RIS, we infer replication fork direction and thus determine the positional relationship between transcription and replication. Surprisingly, head-on transcription occurs frequently at both intergenic and intragenic RIS in asynchronous breast cancer cells. Furthermore, we find that head-on transcription is significantly downregulated in S-phase cells relative to G1-phase cells, especially at R-loop forming sequences. Interestingly, even subtle increases in head-on transcription at RIS have been shown to induce significant DNA damage and replication stress (Hamperl *et al.*, 2017; Nojima et al., 2018), suggesting the downregulation effects we see are likely protective in nature. Collectively, our study identifies pervasive and genic transcription as potential sources of genotoxic head-on collisions, and strongly implicates the existence of a transcriptional regulatory mechanism that functions to silence head-on transcription at RIS during S-phase to preserve genome stability.

Results

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

A multi-layered approach was employed to identify high confidence RIS 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 (Akerman et al., 2020; Foulk et al., 2015). Approximately ~40% of core origin loci are active in any given cell type (Akerman *et al.*, 2020). 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 (Martin *et al.*, 2011). To further filter out false positives, we intersected the remaining loci with an epigenetic signature that predicts binding locations of the origin of replication complex (ORC) with remarkable accuracy (Miotto *et al.*, 2016). This approach yielded 4,572 RIS with a median size of 730 bp.

We validated the identified RISs by assessing their positioning relative to MCF-7 repli-seq replication timing (RT) profiles (Consortium, 2012; Dellino *et al.*, 2013). RT profiles contain an inverted V-apex at sites of replication initiation, and typically apex locations contain one or more bonafide RIS (Dellino *et al.*, 2013). Thus, if our identified RIS loci were true positives, then a high percentage of them should localize within apex regions. Viewing the RIS on a browser track with RT data clearly showed positioning at apex locations in the earliest S-phase fraction (G1b) (Figure 1B, left panel). To assess whether the RIS localized to apexes genome-wide, we assigned an s50 score to each RIS. 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 (Dellino *et al.*, 2013). 68% of total RIS 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 Dnase-seq peaks, demonstrating that our strategy successfully enriched for true RIS (Figure 1B, right panel). To further analyze this subset, RIS were normalized to the median size of 730 bp and centered on a heat map 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 enrichment within the demarcated RIS regions (Figure 1C). Among all RIS, 1,166 localized in intergenic space, 3,030 localized within gene bodies, and 376 spanned gene body termini and adjacent intergenic regions (Figure 1D), in agreement with the distribution of SNS-seq peaks seen in the MCF-7 cell line (Martin *et al.*, 2011). For intergenic RIS, 40% were within 5kb 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 intragenic RIS, we found that 50%, 34%, and 16% localized in this manner (Figure 1E). With an understanding of RIS positioning relative to gene units, we next sought to evaluate local transcription at these sites.

Head-on transcription occurs at intergenic and intragenic RIS

To assess whether head-on transcription occurs at or near RIS and whether it was a feature of genic or pervasive transcription, we separately evaluated transcription at intergenic and intragenic subsets of RIS. We first measured transcription initiation

observed within 3kb of the RIS. To positionally map transcription initiation at these sites, we utilized published data from Mnase-seq (Shimbo et al., 2013), Dnase-seq (Consortium, 2012), RNAP2 ChIP-seq (Consortium, 2012), TBP ChIP-exo (Venters and Pugh, 2013), and TSS-seq (Yamashita et al., 2011). The results show that transcription initiated within a nucleosome-depleted region (NDR) adjacent to both RIS subsets, in agreement with past studies evaluating the chromatin landscape around ORC ChIP-seq and SNS-seq peaks in yeast and murine cells, respectively (Eaton et al., 2010; Foulk *et al.*, 2015) (Figure 2A). Thus, proximal transcription initiation is a feature of the local RIS environment. Due to this conserved organization, we were able to uniformly orient all RIS so that transcription initiation was downstream on a heatmap, enabling positional analysis of transcriptional activity at these loci on a global scale (Figure 2A).

GRO-seq is a highly sensitive nuclear run-on assay capable of mapping genic and pervasive transcription with strand specificity (Core et al., 2008), including unstable and lowly expressed transcripts. To evaluate transcriptional activity at RIS, we first utilized asynchronous GRO-seq data generated in the MCF-7 cell line (Liu et al., 2017). We assessed GRO-seq signal traveling out of the downstream NDR and into the upstream RIS. We called this head-on (HO) transcription because it converges into an emerging replication fork. On both the global and individual locus scale, we found a strong peak of transcription initiating near the border of the RIS adjacent to the NDR and peaking within the center of the RIS. These data reveal that HO transcription is a feature of the local RIS environment (Figure 2A,B). HO transcription was evident at both intergenic and intragenic RIS (Figure 2A,B), demonstrating that it is an intrinsic feature of this RIS

subset and not due to association with gene bodies. Analysis of HO GRO-seq read density at the two subsets of RIS relative to gene GRO-seq read density showed that HO transcription occurs at a similar frequency as highly expressed gene transcription (Figure 2C).

To more systematically evaluate HO transcription at RIS, we next sought to identify the frequency of head-on transcription units (HO TUs). To do this, we utilized MCF-7 NET CAGE-seq data, which identifies TSSs genome-wide with high sensitivity and directional information via cap chemistry and sequencing (Hirabayashi et al., 2019). We defined HO TUs as regions bookended on one end by a HO NET CAGE-seq peak within 1kb of an RIS border, and on the other the RIS summit (Supplemental Figure 3A). We found that 3,357 of the 4,572 RIS contained at least one HO TU (Supplemental Figure 3B). In total, we identified 4,567 HO TUs, as multiple units formed at some RIS. 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 RIS summit (Supplemental Figure 3C). Thus, in agreement with earlier analysis, HO transcription is a feature of a majority of RIS, and occurs within distinct, identifiable units.

We next evaluated whether the HO transcription observed at RIS was pervasive in nature. The intergenic RIS subset was localized completely outside gene bodies and, as such, the transcription occurring at these regions was pervasive. However, the intragenic RIS subset is located within gene bodies, making is unclear if HO

transcription at these sites is generated from normal genic transcription or intragenic pervasive transcription. To distinguish between these two possibilities, we measured transcription initiation and GRO-seq signal at intragenic RIS subset by its distance from the nearest genic TSS. Regardless of TSS distance, we observed enrichment of RNAP2, TBP, and TSS-seq signal within the adjacent NDR, suggesting that transcription initiation at RIS loci is occurring independent of a nearby genic TSS (Supplemental Figure 1A). Furthermore, we found that 31% of intragenic RIS contained HO transcripts that traveled antisense to genic transcription (Supplemental Figure 1B). Finally, we determined whether HO transcripts were overrepresented in the GRO-seq dataset relative to total RNA-seq data from the same study (Liu *et al.*, 2017), which would be relatively depleted of HO transcripts if these were indeed pervasive and thus unstable. We found a significant reduction in total RNA-seq RPKM values for HO transcripts relative to GRO-seq, further suggesting that HO transcription at intragenic RIS is pervasive in nature (Supplemental Figure 1C).

As an orthogonal approach, we interrogated identified HO TUs for pervasive characteristics. Like RIS units, HO TUs had higher GRO-seq RPKM values than RNA-seq (Supplemental Figure 3D). In contrast, active genes had higher RNA-seq RPKM values then GRO-seq, reinforcing the validity of this approach (Supplemental Figure 3D). We next evaluated whether HO TUs associate with different pervasive transcript species, and if so, at what frequency. To do this, 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 (Liu et al., 2022; Whyte et al., 2013). We then categorized HO TUs by whether they overlapped with any of these pervasive transcript classes. We found that 11% of HO TU associations were with PROMPTs, 16% with eRNAs, 18% with asTSSa, 34% with sTSSa, and 21% with transcripts outside these classes (Supplemental Figure 4E). Finally, we observed GRO-seq and RNA-seq values across HO TUs categorized by transcript class association. Interestingly, we found that HO TUs had higher GRO-seq RPKMs across associations, reinforcing that HO TUs are indeed pervasive in nature (Supplemental Figure 3F).

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

Head-on transcription at RIS is markedly downregulated in actively replicating cells

The results from asynchronous MCF-7 cells described above raise the question of how could head-on transcription occur at RIS without negative effects on cellular fitness? We hypothesized that although head-on transcription occurs during the cell cycle, it might be mitigated during genome replication in S-phase. We therefore evaluated HO GRO-seq RPKM distributions at intergenic and intragenic RIS between MCF-7 cells

synchronized in either S-phase or G1-phase (Liu *et al.*, 2017). The results show there is a marked decrease in GRO-seq signal in S-phase cells relative to G1-phase cells across both RIS subsets (Figure 3A,B,C). Importantly, while we observed a small overall decrease in gene transcription between S-phase and G1-phase cells, the magnitude of the HO transcriptional changes at RIS were significantly greater, demonstrating that S-phase downregulation is biased towards RIS (Figure 3D). Moreover, the transcription of genes with proximal upstream RIS is not downregulated in S-phase, demonstrating that the effects seen at RIS are independent of transcriptional buffering that might occur on replicated DNA (Padovan-Merhar et al., 2015; Yunger et al., 2018) (Supplemental Figure 2). These analyses indicate that HO pervasive transcription at RIS is selectively downregulated during S-phase, suggesting that temporally tuned transcriptional regulation at RIS might play a role in genome stability.

We next assessed transcriptional dynamics at HO TUs. In agreement with the previous RIS-based analysis, we found that HO TU transcription was significantly downregulated in S-phase relative to G1-phase cells (Supplemental Figure 4A,B). Differential expression analysis revealed that 1,827 HO TUs are significantly downregulated in S-phase, while only 28 HO TUs are significantly upregulated (Supplemental Figure 4C). Moreover, significant reductions in HO TU transcription levels during S-phase were apparent across transcript class associations, suggesting S-phase suppression is a feature of HO TUs, and not pervasive transcript species broadly speaking (Supplemental Figure 4D,E). Importantly, we found that randomly selected size and

transcriptional activity matched TUs within gene bodies did not show S-phase specific downregulation (Supplemental Figure 5A). Interestingly, pervasive TUs, protein-coding genes, and lincRNAs showed a slight bias towards S-phase downregulation (Supplemental Figure 5B,C,D). 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 S-phase downregulation (Supplemental Figure 5E). Collectively, these findings support the idea that head-on transcription at RIS is specifically suppressed during genome replication.

We also sought to determine if genic transcription through genes containing RISs was downregulated during S-phase. Indeed, RIS-containing genes were downregulated to a greater degree than genes lacking RIS, suggesting that head-on genic transcription is preferentially downregulated during genome replication (Figure 3E). Collectively, the data in Figures 1 through 3 suggest a model in which head-on pervasive and coding transcription occurs during the cell cycle, but is reduced during S-phase, potentially to avoid genotoxic collisions with the replisome.

S-phase downregulation of head-on pervasive transcription is amplified over R-loop forming sequences

If a particular sequence feature amplified the genotoxicity of head-on collisions, one would predict that RIS containing this feature would experience a greater degree of S-phase transcription downregulation. R-loops are three-stranded nucleic acid structures generated when nascent RNA anneals to the DNA template strand during transcription

(Aguilera and Garcia-Muse, 2012). It has been reported that HO transcription-replication collisions generate DNA damage through stabilizing R-loops over C-rich R-loop forming sequences (RLFS) (Hamperl *et al.*, 2017). To address whether S-phase transcription was preferentially downregulated at RLFS, we first identified RLFS annotated by R-loopDB (Jenjaroenpun et al., 2017) within the HO template strand at RIS. RLFS occurred within the HO template strand at 63% of all RIS loci, relative to only 2% of random loci within the template strand of gene bodies (data not shown). This finding suggested that HO transcription at RIS has a predisposition to form R-loops, and enabled us to quantitate differences in transcription between RLFS positive and negative RIS. We observed a sharper loss of S-phase transcription signal at RLFS positive RIS, despite both subsets exhibiting similar levels of transcription in G1-phase (Figure 4A). Moreover, when we compared the temporal RPKM distributions of HO GRO-seq reads at RIS subsets, we found that only RLFS positive RIS RPKMs displayed a significant downward shift in S-phase relative to G1-phase (Figure 4B). Differential expression analysis using a volcano plot revealed that 42% of RLFS high RIS demonstrated significant downregulation based on pre-determined thresholds (see methods), whereas only 22% of RLFS low RIS were significantly downregulated (Figure 4C). However, it is clear from the plot that the vast majority of transcription at both RIS subsets is downregulated. The INO80 chromatin remodeling complex (INO80C) has been shown to prevent R-loop dependent damage during DNA replication in MCF-7 cells (Prendergast et al., 2020). Interestingly, ChIP-seq revealed that INO80C binds non-randomly at MCF-7 RIS, with significantly increased occupancy at the RLFS positive subset, providing a link between a genome protectant and temporally regulated

RIS (Figure 4D,E,F). In sum, this analysis demonstrates that HO transcription is preferentially downregulated at collision sites with high genotoxic potential, supporting the idea that HO RIS transcription is actively regulated across the cell cycle to prevent DNA damage and maintain genome stability.

We lastly evaluated RLFS frequency and positioning within the transcribed strand of HO TUs. We found that 3,476 of the 4,567 HO TUs contained at least one RLFS on the template strand, and that RLFS appeared to localize throughout the HO TU body (Supplemental Figure 6A,B). We next looked at temporal transcriptional changes at HO TUs subset by increasing RLFS density levels. We observed a clear relationship between increasing RLFS density and S-phase downregulation, again suggesting that temporal suppression of HO TUs is likely a mechanism to prevent genotoxic transcription-replication collisions (Supplemental Figure 6C). In aggregate, we propose that HO TUs potentiate damaging collisions with the replisome, and are actively silenced in S-phase by still unknown players to prevent DNA damage (Supplemental Figure 6D).

Discussion

Head-on transcription-replication collisions 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 (Petryk *et al.*, 2016). Our analysis, utilizing multiple published datasets from the MCF-7 breast cancer model, reveals a novel source of head-on transcriptional complexes stemming from pervasive

transcription initiating within an accessible region immediately adjacent to RIS (Akerman *et al.*, 2020; Consortium, 2012; Liu *et al.*, 2017; Martin *et al.*, 2011; Miotto *et al.*, 2016; Venters and Pugh, 2013). Furthermore, our analysis demonstrates that head-on pervasive transcription at these sites is downregulated in S-phase, suggesting that collisions are minimized through a transcriptional regulatory mechanism. In support of the idea that head-on transcription is regulated to maintain genome stability, we find that RLFS-positive RIS, which potentiate highly genotoxic collisions, experience a higher magnitude of regulation than RLFS-negative RIS. Furthermore, the effect sizes we observe in our data likely indicate functional avoidance of genotoxicity. For example, only a 10% increase in HO transcription in an episomal system was shown to generate an 80% loss of the episome (Hamperl *et al.*, 2017). Additionally, an induced 2-fold increase in lncRNA transcription through origins generated severe replication stress in HeLa cells (Nojima *et al.*, 2018). 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 RIS during S-phase.

Although this study is the first in-depth analysis of pervasive transcription dynamics at RIS to our knowledge, the regulation of pervasive transcription has been previously linked to DNA damage prevention in both yeast and mammalian cells (Nojima *et al.*, 2018; Topal et al., 2020). For example, depletion of the positive transcription elongation factor Spt6 in asynchronous HeLa cells was shown to increase PROMPT transcription into origins, R-loop formation, DNA damage, and replication stress (Nojima *et al.*, 2018).

Additionally, co-depletion of the chromatin remodeling complex INO80C and the TBP antagonist MOT1 led to increased head-on transcription and replication stress-dependent DNA breaks at yeast origins (Topal *et al.*, 2020). While these findings strongly suggested that suppression of pervasive transcription was functionally preventing transcription-replication collisions, they did not define if transcriptional silencing by these factors was occurring during genome replication. Our results complement these studies, which support the idea that pervasive transcription at RIS potentiates collisions and must be actively regulated during S-phase.

Acknowledgements

This work was supported by NIH grants R01 GM074701 to M.F.C., and the TRDRP 2019B Predoctoral fellowship award T30DT0906 to M.K.

Figures

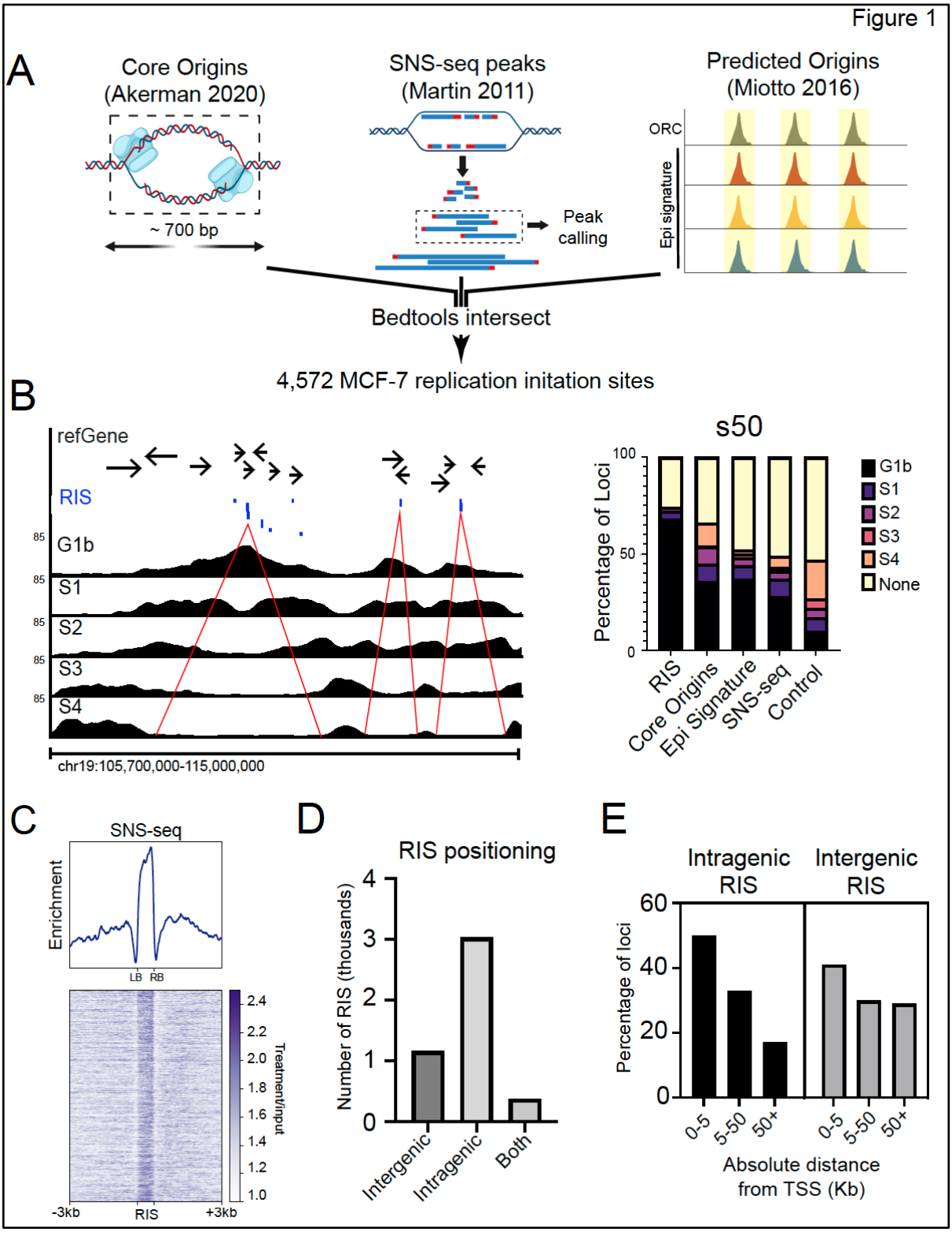


Figure 1
A
Core Origins
(Akerman 2020)
~ 700 bp
SNS-seq peaks
(Martin 2011)
Peak
calling
Predicted Origins
(Miotto 2016)
ORC
Epi signature
Bedtools intersect
4,572 MCF-7 replication initation sites
B
refGene
RIS
G1b
S1
S2
S3
S4
chr19:105,700,000-115,000,000
s50
Percentage of Loci
100
50
0
G1b
S1
S2
S3
S4
None
RIS
Core Origins
Epi Signature
SNS-seq
Control
C
SNS-seq
Enrichment
LB
RB
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
Treatment/Input
-3kb
RIS
+3kb
D
RIS positioning
Number of RIS (thousands)
4
3
2
1
0
Intergenic
Intragenic
Both
E
Intragenic
RIS
Intergenic
RIS
Percentage of loci
60
40
20
0
0-5
5-50
50+
0-5
5-50
50+
Absolute distance
from TSS (Kb)

Figure 2

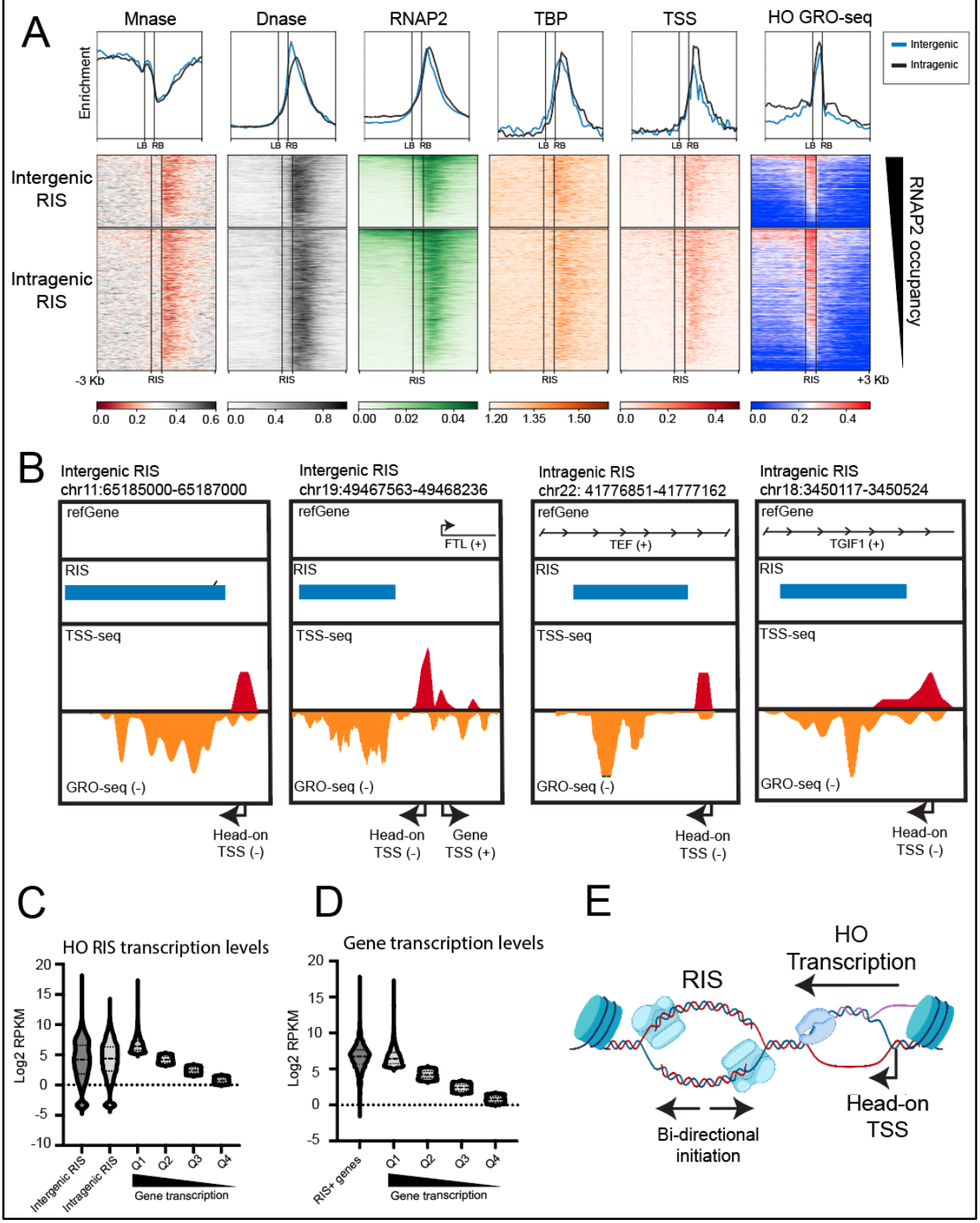

A
Mnase
Dnase
RNAP2
TBP
TSS
HO GRO-seq
Intergenic
Intragenic
Enrichment
LB RB
Intergenic RIS
Intragenic RIS
RNAP2 occupancy
-3 Kb
RIS
+3 Kb
B
Intergenic RIS
chr11:65185000-65187000
Intergenic RIS
chr19:49467563-49468236
Intragenic RIS
chr22: 41776851-41777162
Intragenic RIS
chr18:3450117-3450524
refGene
FTL (+)
TEF (+)
TGIF1 (+)
RIS
TSS-seq
GRO-seq (-)
Head-on
TSS (-)
Gene
TSS (+)
C
HO RIS transcription levels
Log2 RPKM
Intergenic RIS
Intragenic RIS
Q1
Q2
Q3
Q4
Gene transcription
D
Gene transcription levels
RIS+ genes
E
HO
Transcription
RIS
Head-on
TSS
Bi-directional
initiation

Figure 3

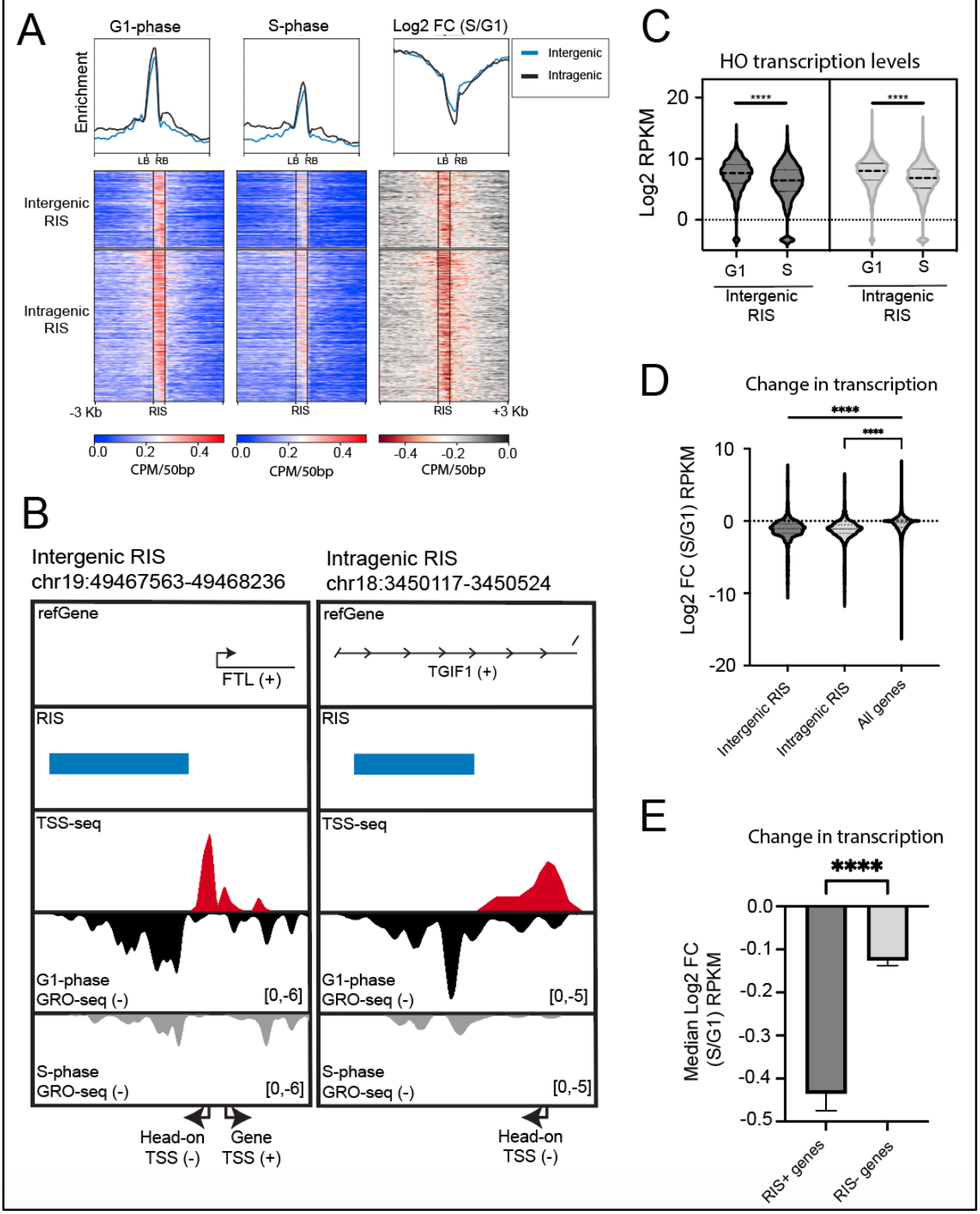

A
G1-phase
S-phase
Log2 FC (S/G1)
Intergenic
Intragenic
Enrichment
LB RB
Intergenic RIS
Intragenic RIS
-3 Kb
RIS
+3 Kb
0.0 0.2 0.4
CPM/50bp
-0.4 -0.2 0.0
B
Intergenic RIS
chr19:49467563-49468236
Intragenic RIS
chr18:3450117-3450524
refGene
FTL (+)
TGIF1 (+)
RIS
TSS-seq
G1-phase
GRO-seq (-)
[0,-6]
[0,-5]
S-phase
GRO-seq (-)
Head-on
TSS (-)
Gene
TSS (+)
C
HO transcription levels
Log2 RPKM
20
10
0
****
G1
S
Intergenic RIS
Intragenic RIS
D
Change in transcription
Log2 FC (S/G1) RPKM
10
0
-10
-20
Intergenic RIS
Intragenic RIS
All genes
E
Change in transcription
Median Log2 FC (S/G1) RPKM
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
RIS+ genes
RIS- genes

Figure 4

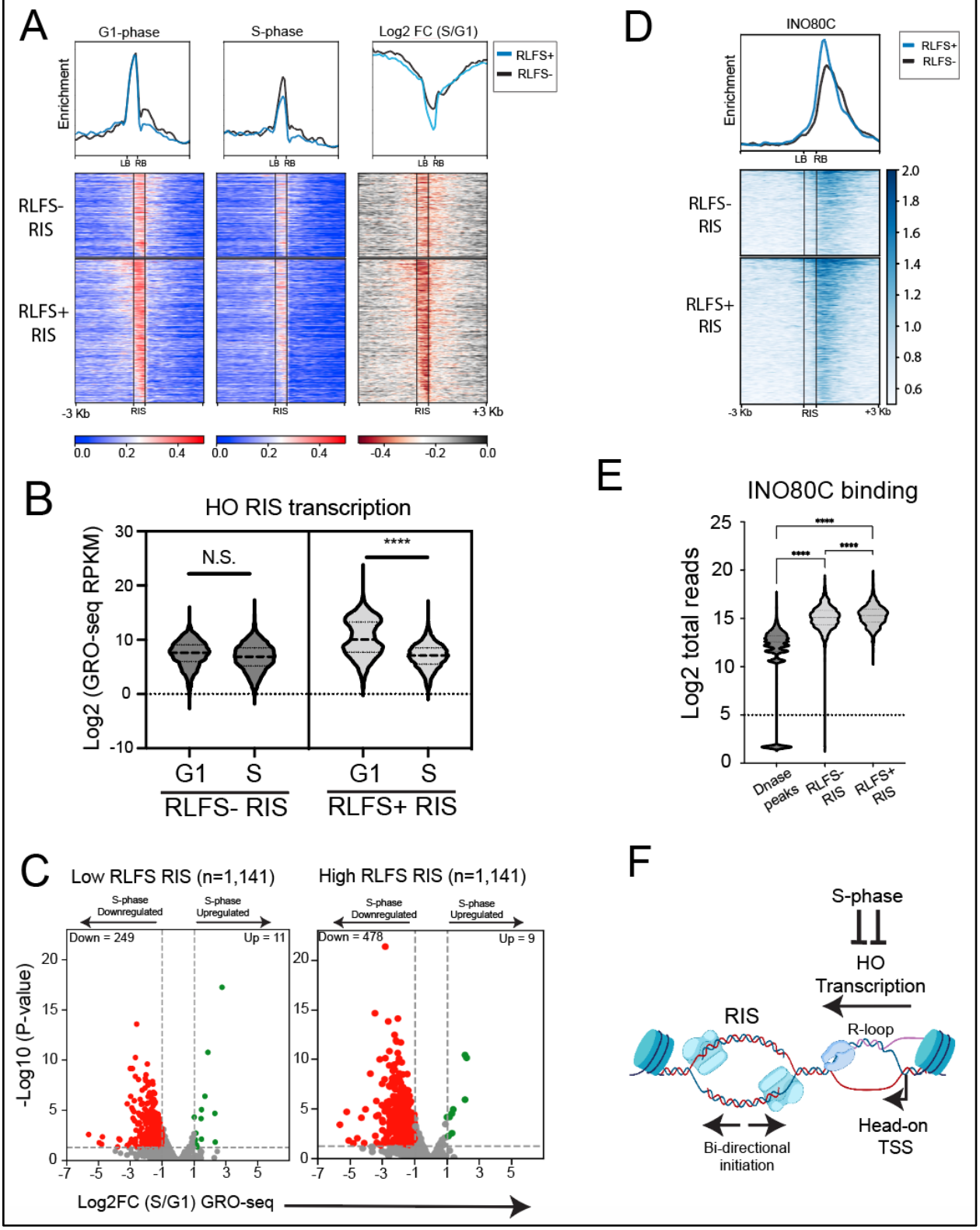

A
G1-phase
S-phase
Log2 FC (S/G1)
Enrichment
RLFS+
RLFS-
LB RB
RLFS- RIS
RLFS+ RIS
-3 Kb
RIS
+3 Kb
0.0 0.2 0.4
-0.4 -0.2 0.0
D
INO80C
2.0 1.8 1.6 1.4 1.2 1.0 0.8 0.6
B
HO RIS transcription
Log2 (GRO-seq RPKM)
N.S.
****
G1 S G1 S
RLFS- RIS
RLFS+ RIS
E
INO80C binding
Log2 total reads
Dnase peaks
RLFS- RIS
RLFS+ RIS
C
Low RLFS RIS (n=1,141)
High RLFS RIS (n=1,141)
S-phase Downregulated
S-phase Upregulated
Down = 249
Up = 11
Down = 478
Up = 9
-Log10 (P-value)
Log2FC (S/G1) GRO-seq
F
S-phase
HO Transcription
RIS
R-loop
Bi-directional initiation
Head-on TSS

Figure Legends

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

A. Schematic of the strategy used to identify MCF-7 RIS. B. (Left) Browser track showing RIS (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 RIS, benchmark, and control datasets. C. Average profile and heatmap of MCF-7 SNS-seq Poisson enrichment at distance normalized RIS loci. D. Bar graph showing RIS frequency by position relative to gene bodies. E. Bar graphs showing RIS frequency by absolute distance relative to the nearest protein-coding TSS.

Figure 2: Head-on transcription occurs at intergenic and intragenic RIS

A. Average profiles and heatmaps of MCF-7 Mnase-seq, Dnase-seq, RNAP2 ChIP-seq, and TBP ChIP-exo Poisson enrichment, and TSS-seq and HO GRO-seq counts per million at distance normalized RIS loci. Black lines align with the left and right boundaries (LB and RB) of the RIS region. B. Browser track examples of transcription at intragenic and intergenic RIS. C. Violin plot showing the distribution of RPKM values for HO GRO-seq reads over subset RIS 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 RIS, and genes split into quartiles by transcription levels. E. Cartoon model showing positional relationship between HO transcription and replication initiation at RIS.

Figure 3: Head-on transcription at RIS is markedly downregulated in actively replicating tumor cells

A. Average profiles and heatmaps of head-on (HO) GRO-seq signal in counts per million at distance normalized RIS loci (Left and Middle panels). GRO-seq from G1-phase cells (Left). GRO-seq from S-phase cells (Middle). Average profiles and heatmaps of the log2 fold change between S-phase and G1-phase CPM values (Right panel). Black lines align with the left and right boundaries of the RIS 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 RIS regions from G1-phase and S-phase cells D. Violin plot comparing the distributions of the fold changes in either HO GRO-seq reads or genic GRO-seq reads between S-phase and G1-phase cells across intergenic RIS, intragenic RIS, and all protein coding genes. E. Bar graph showing the median fold change in genic transcription across genes with and without internal RIS.

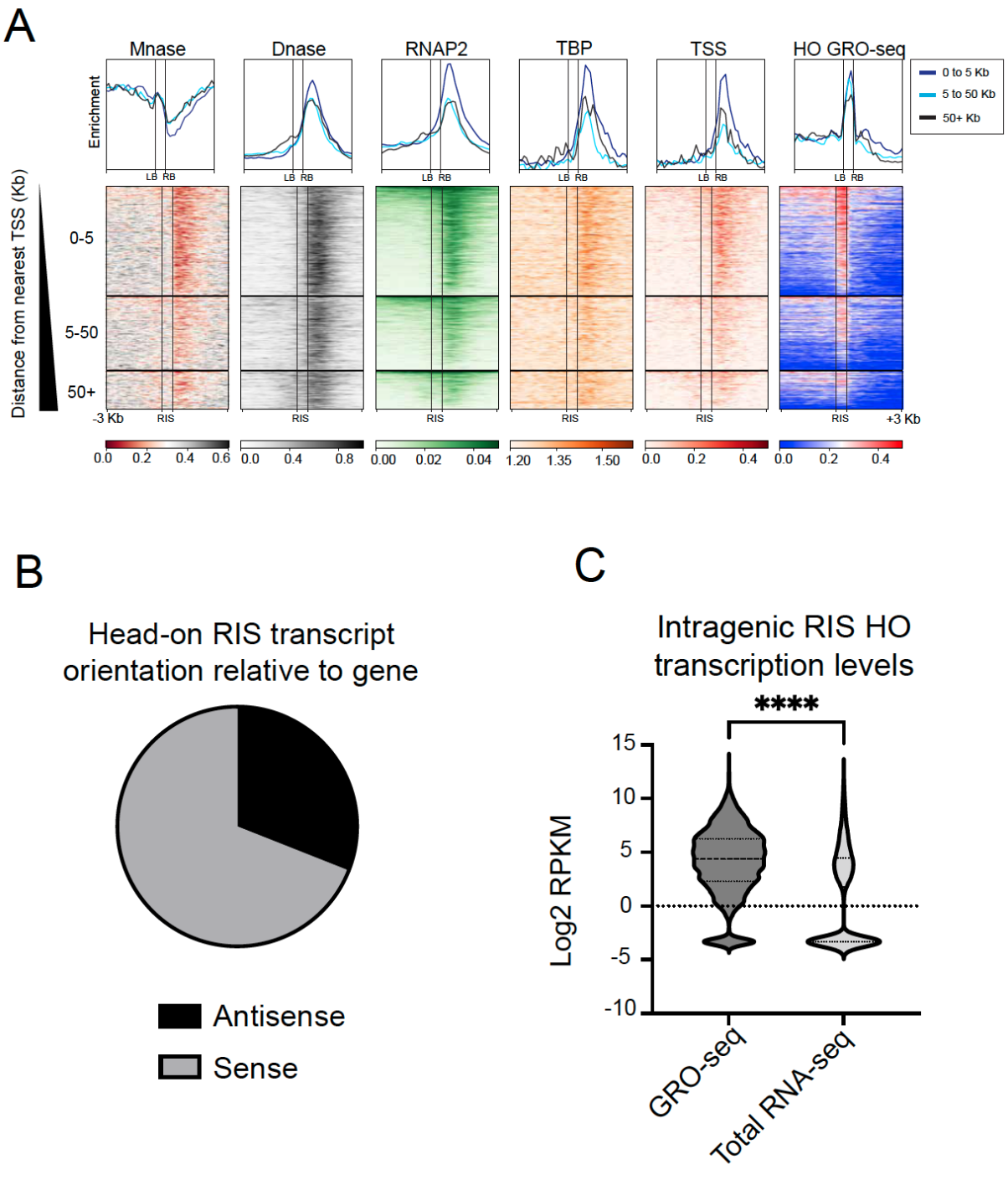
Figure 4: S-phase downregulation of head-on pervasive transcription is amplified over R-loop forming sequences

A. Average profiles and heatmaps of head-on (HO) GRO-seq signal in counts per million at distance normalized RIS loci. (Left) GRO-seq from G1-phase cells. (Middle) GRO-seq from S-phase cells. (Right) Average profiles and heatmaps of the log2 fold change between S-phase and G1-phase CPM values. Black lines align with the left and right boundaries of the RIS region. B. Violin plot comparing the distribution of RPKM values for GRO-seq reads in the HO orientation across RIS with and without RLFS in

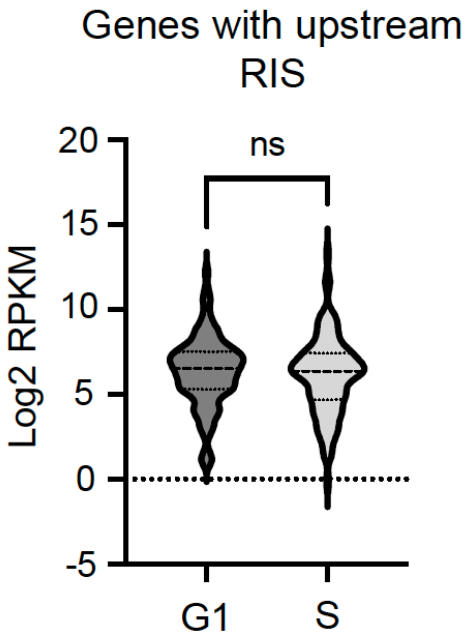
the HO transcript template strand from G1-phase and S-phase cells. C. Volcano plots of a differential expression analysis of HO RPKMs within RIS regions between S-phase and G1-phase cells subset by top 25% or bottom 25% RLFS density. D. Average profile and heatmap of INO80C ChIP-seq Poisson enrichment at distance normalized RIS loci. Black lines align with the left and right boundaries (LB and RB) of the RIS region. E. Violin plot comparing the distribution of total INO80C ChIP-seq reads at randomly selected Dnase-seq peaks, RLFS- RIS, or RLFS+ RIS. F. Cartoon model showing positional relationship between HO transcription, replication initiation at RIS, and cell-cycle specific HO transcription regulation.

Supplemental Figures

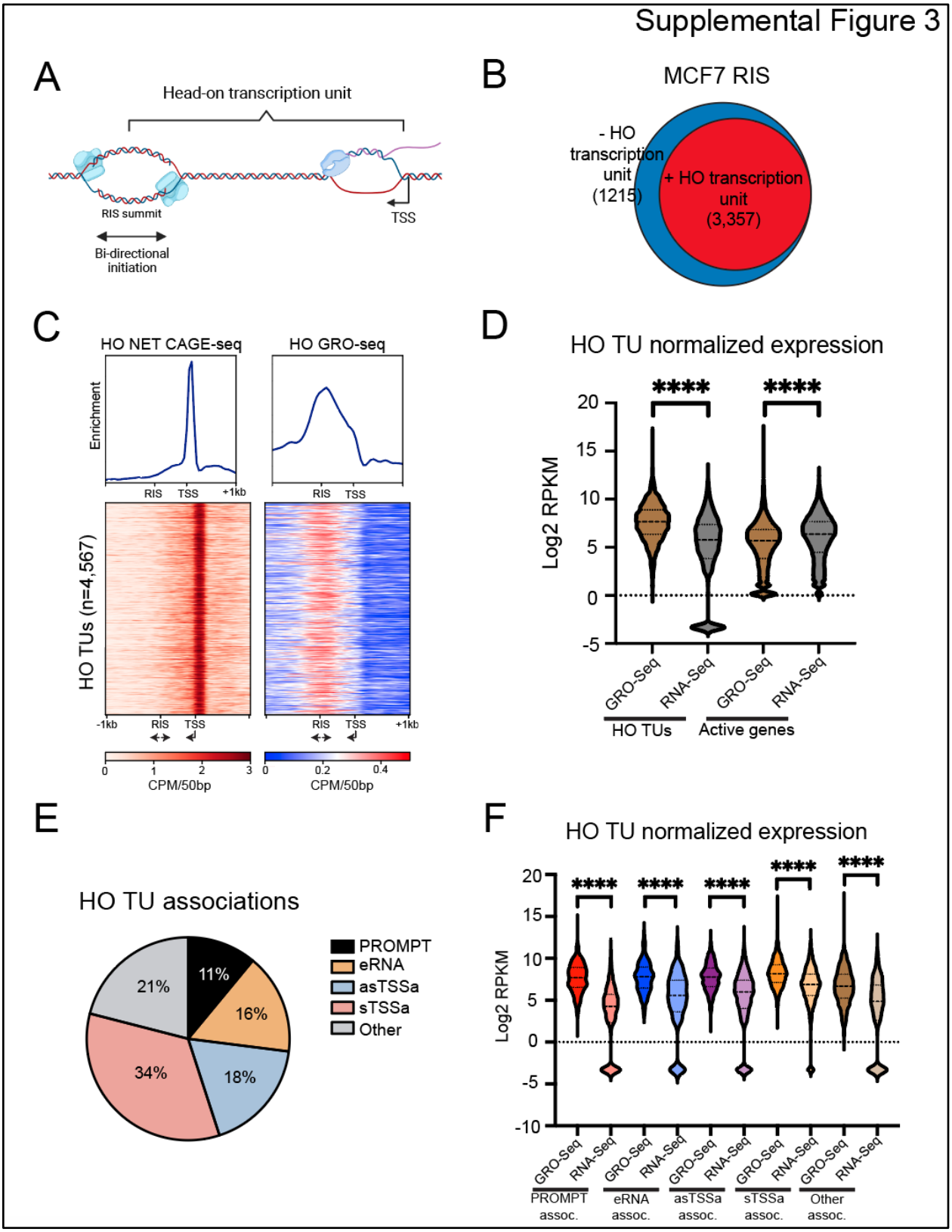
Supplemental Figure 1



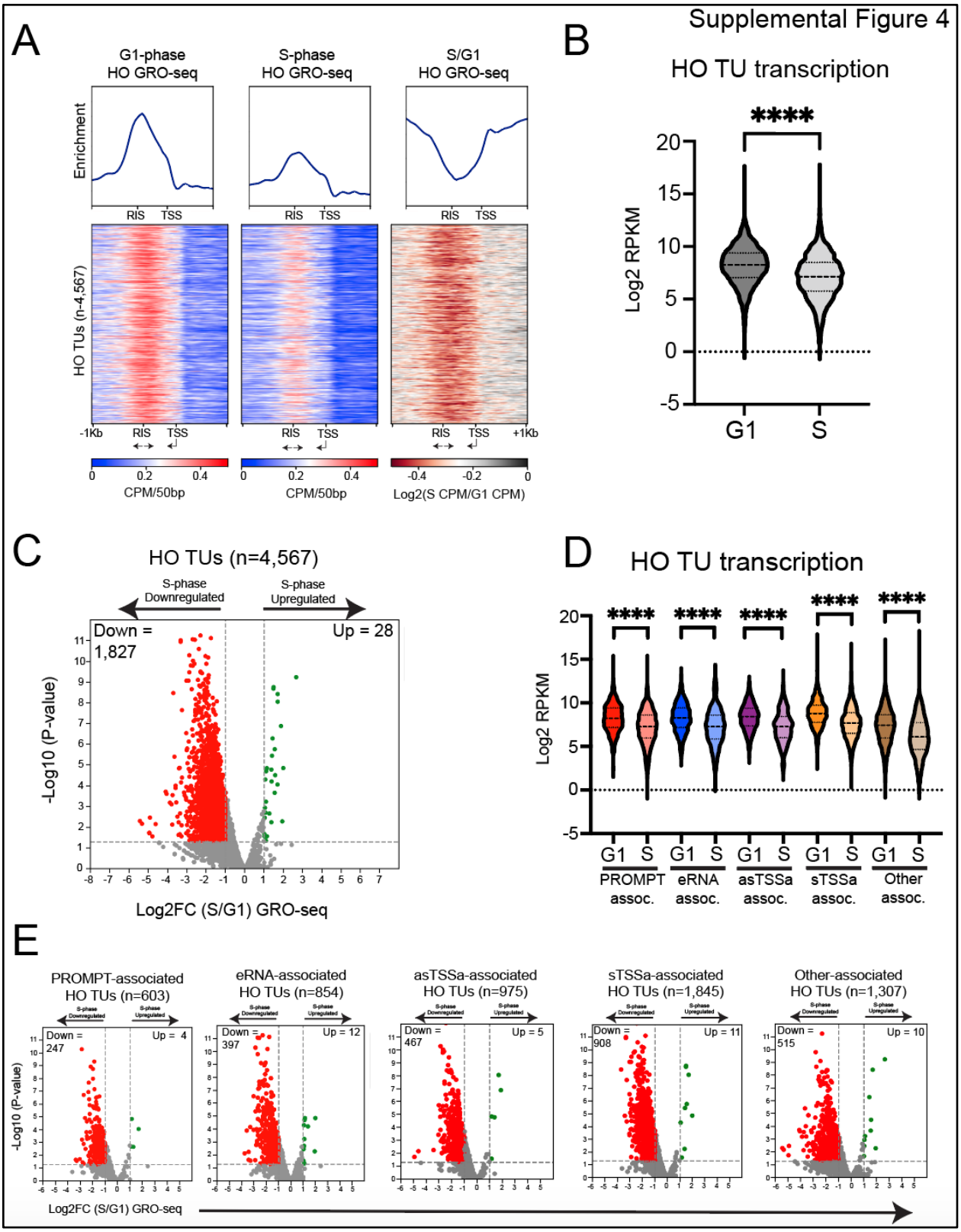
Supplemental Figure 2



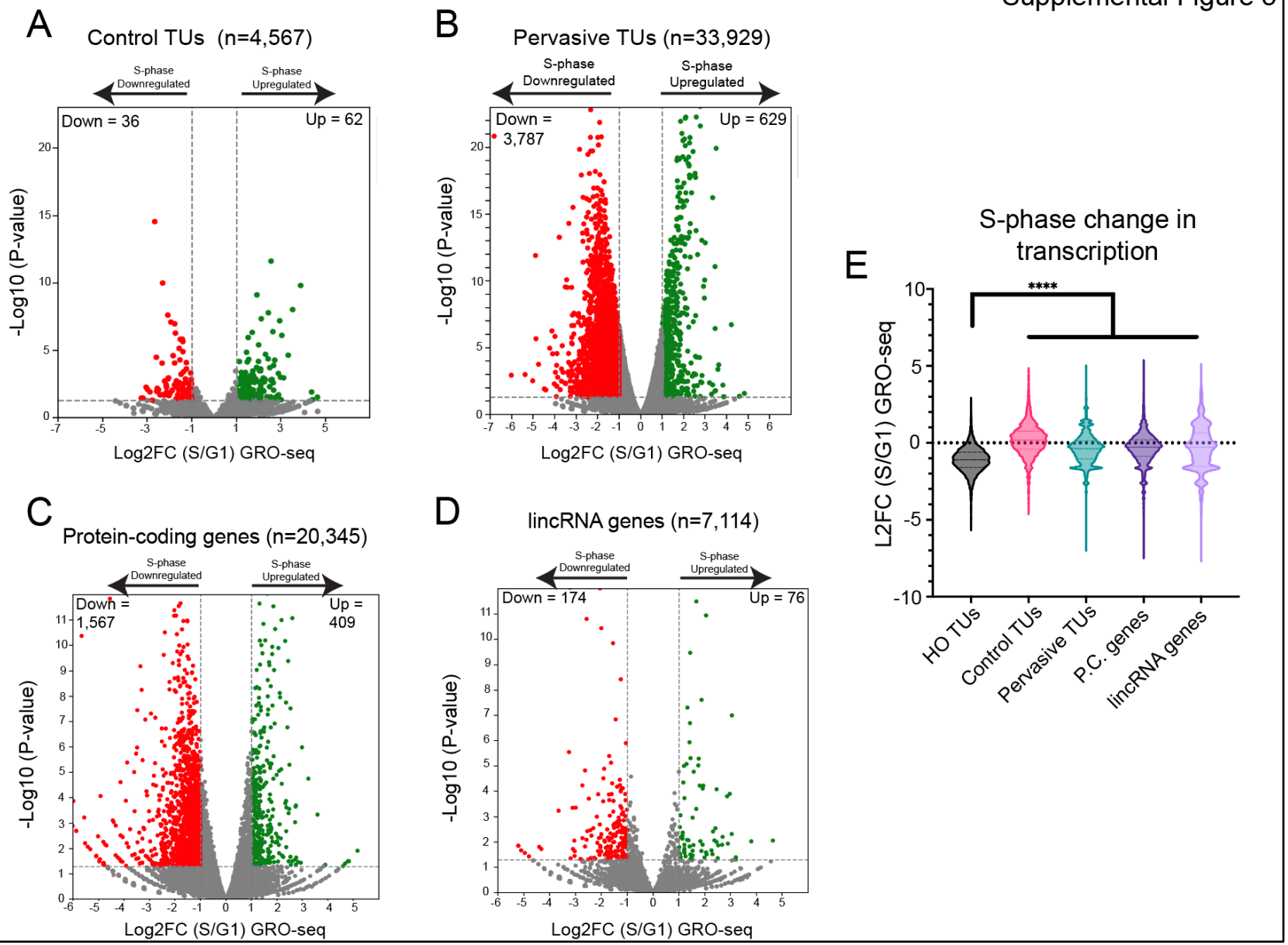
Supplemental Figure 3



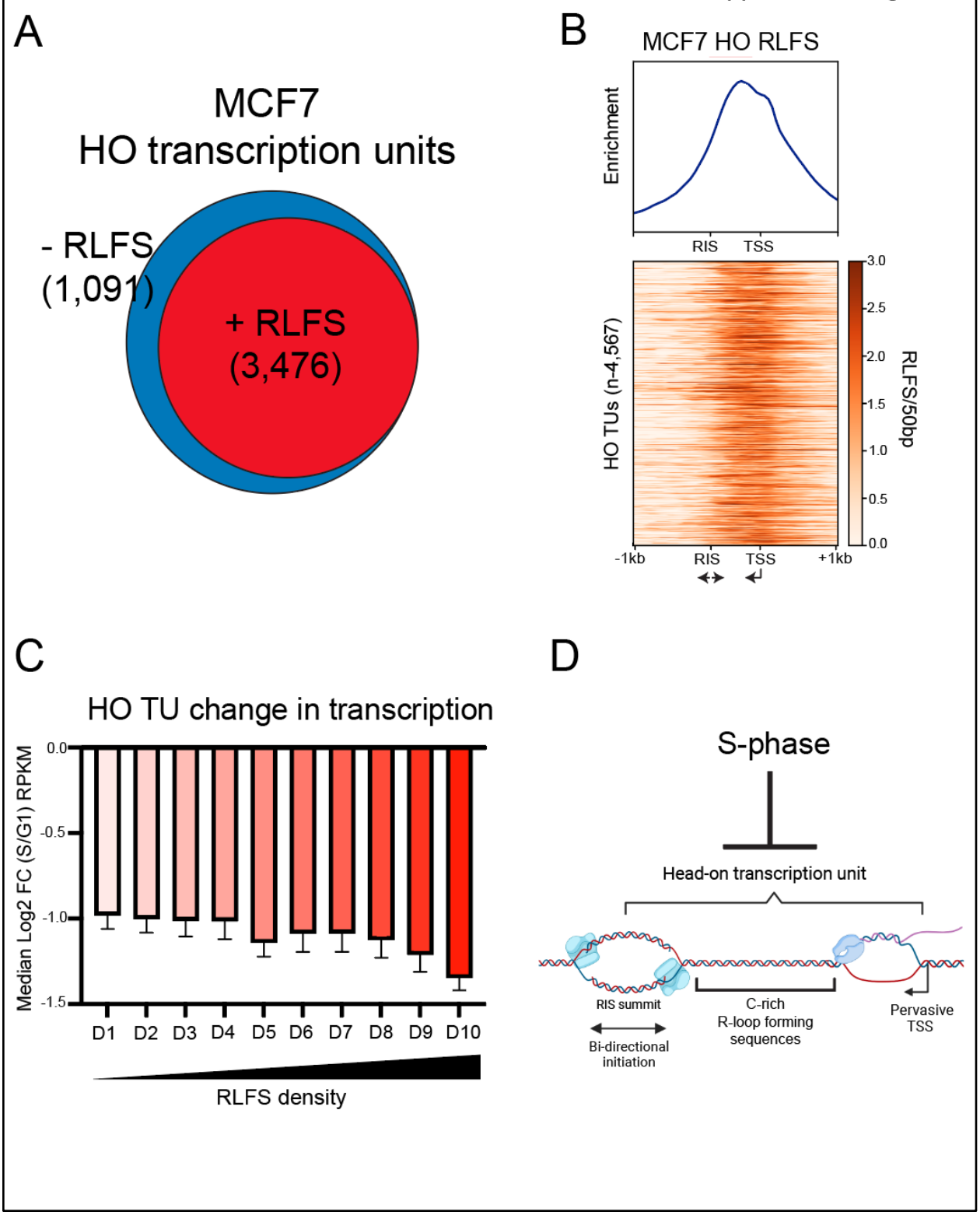
Supplemental Figure 4



Supplemental Figure 5



Supplemental Figure 6



Supplemental Figure Legends

Supplemental Figure 1: HO transcription at intragenic RIS is pervasive in nature.

A. Average profiles and heatmaps of MCF-7 Mnase-seq, Dnase-seq, RNAP2 ChIP-seq, and TBP ChIP-exo Poisson enrichment, and TSS-seq and HO GRO-seq in counts per million within 50bp bins centered on distance normalized RIS regions demarcated by a left boundary (LB) and right boundary (RB). RIS are partitioned into subsets of increasing genomic distance from the nearest protein-coding TSS. B. Pie chart showing the distribution of intragenic RIS with either antisense or sense HO transcriptional activity relative to gene transcription. C. Violin plot showing the distribution of log transformed HO GRO-seq or HO total RNA-seq RPKM values at intragenic RIS.

Supplemental Figure 2: Observed temporal changes in transcriptional activity at RIS occur independently of post-replication transcriptional buffering.

Violin plot comparing the distribution of RPKM values for genic GRO-seq reads over gene bodies with upstream RIS from G1-phase and S-phase cells.

Supplemental Figure 3: HO TUs are a feature of RIS and are pervasive in nature.

A. Graphic representation of a Head-on transcription unit (HO TU). B. Diagram of total RIS demarcated by the presence or absence of at least one HO TU. C. Average profiles and heatmaps of HO CAGE-seq and HO GRO-seq (asynchronous) in counts per million within 50bp bins centered on distance normalized HO TUs demarcated by the TSS and RIS summit. D. Violin plot showing the distribution of HO TU or active gene RPKMs from either the GRO-seq or RNA-seq assay. E. Pie chart showing the percentage of HO

TU associations with a given pervasive TU species. F. Violin plot showing the distribution of HO TU RPKMs subset by pervasive TU association from either the GRO-seq or RNA-seq assay.

Supplemental Figure 4: HO TUs are suppressed during S-phase.

A. Average profiles and heatmaps of head-on (HO) GRO-seq signal in counts per million within 50bp bins at distance normalized HO TUs (Left and Middle panels). GRO-seq from G1-phase cells (Left). GRO-seq from S-phase cells (Middle). Average profiles and heatmaps of the log2 fold change between S-phase and G1-phase CPM values (Right panel). B. Violin plots of HO TU RPKM distributions in G1 and S-phase synchronized cells. C. Volcano plot showing the differential expression of HO TUs between S-phase and G1-phase cells. D. Violin plots of HO TU RPKM distributions in G1 and S-phase synchronized cells subset by pervasive TU association. E. Volcano plots showing the differential expression of HO TUs between S-phase and G1-phase cells subset by pervasive TU association.

Supplemental Figure 5: Observed temporal changes in HO TU transcription are HO TU-specific.

A .Volcano plot showing the differential expression of Control TUs between S-phase and G1-phase cells. B. Volcano plot showing the differential expression of pervasive TUs between S-phase and G1-phase cells. C. Volcano plot showing the differential expression of protein-coding genes between S-phase and G1-phase cells. D. Volcano plot showing the differential expression of lincRNAs between S-phase and G1-phase

cells. E. Violin plots showing the S-phase verse G1-phase GRO-seq log2 fold change distribution of HO TUs and control datasets.

Supplemental Figure 6: HO TUs contain R-loop forming sequences and S-phase regulation is linked to RLFS density.

A. Diagram of total HO TUs demarcated by the presence or absence of at least one RLFS in the template strand. B. Average profile and heatmap of RLFS frequency on template strand within 50bp bins at distance normalized HO TUs. C. Bar chart showing the median log2 fold change (with 95% confidence interval) in S-phase verse G1-phase GRO-seq RPKMs at HO TUs subset by RLFS density. D. Graphic depicting an HO TU enriched in RLFS on the template strand and its temporal regulation.

Materials and Methods

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
ACTR5/Arp5 antibody	Proteintech	Cat# 21505-1-AP, RRID:1234
Deposited Data		
Core origin coordinate file	Akerman et al. 2020	NCBI Gene Expression Omnibus (GEO): GSE128477
MCF-7 SNS-seq	Martin et al. 2011	NCBI Gene Expression Omnibus (GEO): GSE28911
MCF-7 Dnase-seq	John Stamatoyannopoulos, UW	ENCODE: doi:10.17989/ENCSR000EPH
MCF-7 H3K4me2 ChIP-seq	Bradley Bernstein, Broad	ENCODE: doi:10.17989/ENCSR875KOJ
MCF-7 H3K27ac ChIP-seq	Bradley Bernstein, Broad	ENCODE: doi:10.17989/ENCSR752UOD
MCF-7 Repli-seq S1	John Stamatoyannopoulos, UW	ENCODE: doi:10.17989/ENCSR727ZRP
MCF-7 Repli-seq S2	John Stamatoyannopoulos, UW	ENCODE: doi:10.17989/ENCSR170QBY
MCF-7 Repli-seq S3	John Stamatoyannopoulos, UW	ENCODE: doi:10.17989/ENCSR404GFT

MCF-7 Repli-seq S4	**John Stamatoyannopoulos, UW**	**ENCODE: doi:10.17989/ENCSR831UBH**
EJ3 Ini-seq	**Langley et al. 2016**	**European Nucleotide Archive (ENA): PRJEB12207**
K562 ORC1 ChIP-seq	**Miotto et al. 2016**	**NCBI Gene Expression Omnibus (GEO): GSE70165**
MCF-7 H2A.Z ChIP-seq	**Bradley Bernstein, Broad**	**ENCODE: doi:10.17989/ENCSR057MWG**
HaCat G4 ChIP-seq	**Hansel-Hersch et al. 2016**	**NCBI Gene Expression Omnibus (GEO): GSE76688**
MCF-7 RNAP2 ChIP-seq	**Vishwanath Iyer, UTA**	**ENCODE: doi:10.17989/ENCSR000DMT**
MCF-7 TBP ChIP-exo	**Venters et al. 2013**	**NCI read archive: SRA067908**
MCF-7 TSS-seq	**Yamashita et al. 2011**	**NCI read archive: SRA003625**
MCF-7 GRO-seq	**Liu et al. 2017**	**NCBI Gene Expression Omnibus (GEO): GSE94479**
MCF-7 RNA-seq	**Liu et al. 2017**	**NCBI Gene Expression Omnibus (GEO): GSE94479**
MCF-7 NET CAGE-seq	**Hirabayashi et al. 2019**	**NCBI Gene Expression Omnibus (GEO): GSE118075**
R-loop forming sequences	**Jenjaroenpun et al. 2017**	**http://rloop.bii.a-star.edu.sg/**
Experimental Models: Cell Lines		
MCF-7	**ATCC**	**HTB-22**
Software and Algorithms		

Bedtools	**Quinlan and Hall 2010**	
Samtools	**Li et al., 2009**	
Tophat2	**Kim et al., 2013**	
MACS2	**Zhang et al., 2008**	
Bowtie2	**Langmead et al, 2009**	
Deeptools	**Ramirez er al. 2014**	
HOMER	**Heinz et al., 2010**	
ROSE	**Whyte et al. 2013**	

EXPERIMENTAL PROCEDURES

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 (Langmead et al., 2009) to generate bam alignment files. All bam files were then processed with samtools (Li et al., 2009) so that duplicates were removed, and low-quality reads were filtered out. MACS2 peakcall (Zhang et al., 2008) was then used to generate read normalized treatment and background bedgraph files from IP and input controls respectively. MACS2 bdgcmp (Zhang *et al.*, 2008) 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 (Consortium, 2012; Kent et al., 2010) for downstream analysis using the python deeptools software suite (Ramirez et al., 2014).

RIS identification

Core origin summits (Akerman *et al.*, 2020), MCF-7 SNS-seq peaks (Martin *et al.*, 2011), and Dnase-seq peaks from loci containing overlapping peaks of MCF-7 H3K27ac ChIP-seq, MCF-7 H3K4me2 ChIP-seq, and MCF-7 Dnase-seq were extended 1 kb in each direction using bedtools slop (Consortium, 2012; Quinlan and Hall, 2010). These extended peaks were then intersected using bedtools intersect (Quinlan and Hall, 2010). Intersected core origin coordinates were used to represent RIS.

RIS validation

Samtools bedcov (Li *et al.*, 2009) was used to map reads from MCF-7 replication timing datasets (Repli-seq) (Consortium, 2012) to RIS 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 1kb in each direction for mapping using bedtools slop (Quinlan and Hall, 2010). For RIS and core origins, the center of all coordinate locations were taken and extended 1kb in each direction for mapping using bedtools slop. 4,572 random Dnase-seq peaks were selected through using bedtools shuffle (Quinlan and Hall, 2010) 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 (Dellino *et al.*, 2013).

RIS global visualization

For heatmap and average profile generation of RIS on a global scale, RIS 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 3kb upstream and downstream of the normalized region (demarcated by a left and right boundary, 'LB' and 'RB'), divided into 50bp bins using the python deeptools computeMatrix function (Ramirez *et al.*, 2014). The matrix was then sorted by largest to smallest RIS region length using the python deeptools plotHeatmap function (Ramirez *et al.*, 2014). An SNS-seq Poisson enrichment bigwig file was then overlaid onto the matrix via the computeMatrix and plotHeatmap functions.

RIS sub-setting by intragenic or intergenic status

Intragenic RIS were identified by using bedtools intersect to find RIS entirely confined within protein-coding gene body termini as annotated from the GENCODE database (Frankish et al., 2019). Intergenic RIS were identified by using bedtools subtract (Quinlan and Hall, 2010) to identify the remaining RIS. If RIS both overlapped gene body regions and adjacent intergenic regions, they were categorized as 'both' and removed from further analysis.

RIS sub-setting by TSS distance

The HOMER annotatePeaks function (Heinz et al., 2010) was used to determine the distance from the nearest protein-coding TSS for each RIS location based off the RIS center coordinate. RIS were then binned by the calculated absolute distance.

Determination of genes with upstream RIS

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

Orienting RIS to proximal transcription initiation events

RIS were uniformly aligned to their proximal NDR region using the python deeptools computeMatrix function (Ramirez *et al.*, 2014) and the processed Dnase-seq bigwig file (ENCODE), with the NDR being oriented downstream of the RIS region on the matrix. Poisson enrichment scores from the generated TBP ChIP-exo bigwig file, RNAP2 ChIP-seq bigwig file, and TSS-seq bigwig file in counts per million were then overlaid onto this aligned matrix using the python deeptools computeMatrix and plotHeatmap functions (Ramirez *et al.*, 2014). All analyses of GRO-seq signal utilize this aligned matrix.

Head-on transcription unit (HO TU) identification

Directional NET CAGE-seq peaks (Hirabayashi et al. 2019) were intersected with regions delimited by a RIS center and 1kb downstream of the RIS border proximal to the NDR using bedtools intersect. Minus strand NET CAGE-seq peaks were intersected with RIS that formed a downstream NDR, and plus strand NET CAGE-seq peaks were intersected with RIS that formed an upstream NDR. Intersected peaks were labeled HO TU TSS, and the cognate RIS center point represented the HO TU terminus. Some RIS contained multiple HO TUs due to multiple NET CAGE-seq peaks intersecting with the demarcated RIS region.

GRO-seq raw data processing

Raw fastq files from (Liu *et al.*, 2017) were mapped to the hg19 genome with tophat2 (Kim et al., 2013) to produce bam alignment files. Duplicates and low quality reads were removed from bam files via samtools (Li *et al.*, 2009). Replicate bam files were merged for downstream analysis using samtools merge (Li *et al.*, 2009). 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 (Ramirez *et al.*, 2014). 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 (Li *et al.*, 2009), 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 5kb 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 (Whyte *et al.*, 2013). This yielded 11,564 total eRNAs. Transcripts were labeled as asTSSa if they overlapped with TSS plus 500bp 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 20-90bp in order to enrich for short pervasive transcripts, yielding 51,492 transcripts. Transcripts were labeled as sTSSa if they overlapped with TSS plus 500bp 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 RIS, GRO-seq stranded bigwig files were directionally mapped to RIS 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 RIS matrix as was previously described, using a 150bp smoothing length (Ramirez *et al.*, 2014). After mapping stranded bigwig files to the directionally subset RIS, the matrices were combined via the deeptools computeMatrix Rbind function for visualization of directional GRO-seq signal across all RIS (Ramirez *et al.*, 2014).

To observe differences between directional GRO-seq signal at RIS 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 off mapped reads from a S2 Drosophila spike-in. Normalized bigwig files could then be mapped as previously described to observe relative signal in counts per million at RIS. To assess log2 fold change signal at RIS, 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 (Ramirez *et al.*, 2014). Bins with values of 0 were replaced with 0.1 for this analysis. These bigwig files could then be directionally mapped onto RIS 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 (Li et al., 2019).

RPKM calculations

Merged and QC'd bam files generated from fastq files from (Liu *et al.*, 2017) as previously described were converted to sam files, separated by strand, reconverted to bam files, and indexed using samtools (Li *et al.*, 2009). To find HO RPKMs, samtools bedcov was used to map reads from stranded bam files directionally onto RIS regions subset by location of the accessible region on the plus strand. Subsequent files containing head-on mapped read information for each RIS subset were then

concatenated. Mapped reads within RIS 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 features

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. The merged RLFS coordinate file for the hg19 genome was downloaded and separated by strand. Concomitantly, RIS were subset by accessible region location based on the plus strand as was done in prior analyses. To identify RIS that contained RLFS in the head-on transcription template strand, bedtools intersect was used to find subset RIS that overlapped with the directionally appropriate stranded RLFS file. Resulting files were then concatenated. The same pipeline was used to assess RLFS presence within the template strand of HO TUs.

To determine RLFS high and low RIS, a bedgraph file describing RLFS frequency per 50bp 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 along with RIS coordinates for

python deeptools analysis. Output files describing RLFS density within individual RIS units were rank-ordered and the bottom 25% and top 25% loci were selected for low and high groups respectively.

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

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 .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 (760bp). 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.

MCF-7 cell culture

MCF-7 cells were cultured on TC qualified plates in media containing DMEM/F12 (1:1) (Thermo Fisher, 11320-033), supplemented with 10% fetal bovine serum.

INO80C ChIP-seq

ChIP-seq was performed in MCF-7 cells as was done in (Xue et al., 2017), using an antibody against the INO80C subunit ACTR5 (ProteinTech Cat# 21505-1-AP). Generated fastq files were processed as described previously to produce bam alignment files and bigwig files for downstream analysis.

INO80C RIS occupancy analysis

QC'd bam files generated from INO80C ChIP-seq fastq files as previously described were indexed using samtools (Li *et al.*, 2009). Samtools bedcov was used to map reads from bam files onto regions that extended 1 kb from the RIS boundary into the NDR. A random Dnase-seq peak file (described previously) was uniformly extended 500bp in each direction to generate 1kb control regions.

Graphics generation

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

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.

References

Aguilera, A., and Garcia-Muse, T. (2012). R loops: from transcription byproducts to threats to genome stability. Mol Cell *46*, 115-124. 10.1016/j.molcel.2012.04.009.

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. 10.1038/s41467-020-18527-0.

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

Candelli, T., Gros, J., and Libri, D. (2018). Pervasive transcription fine-tunes replication origin activity. Elife *7*. 10.7554/eLife.40802.

Chen, Y.H., Keegan, S., Kahli, M., Tonzi, P., Fenyo, 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. 10.1038/s41594-018-0171-0.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74. 10.1038/nature11247.

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. 10.1126/science.1162228.

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. 10.1101/gr.142331.112.

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. 10.1101/gad.1913210.

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. 10.1101/gr.183848.114.

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. 10.1093/nar/gky955.

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 e719. 10.1016/j.cell.2017.07.043.

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. 10.1371/journal.pgen.1003569.

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. 10.1016/j.molcel.2010.05.004.

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. 10.1038/s41588-019-0485-9.

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. 10.1074/jbc.M113.492504.

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. 10.1093/nar/gkw1054.

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. 10.1093/bioinformatics/btq351.

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. 10.1186/gb-2013-14-4-r36.

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 e718. 10.1016/j.cell.2017.07.044.

Langley, A.R., Graf, 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. 10.1093/nar/gkw760.

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. 10.1186/gb-2009-10-3-r25.

Li, D., Hsu, S., Purushotham, D., Sears, R.L., and Wang, T. (2019). WashU Epigenome Browser update 2019. Nucleic Acids Res *47*, W158-W165. 10.1093/nar/gkz348.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079. 10.1093/bioinformatics/btp352.

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

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. 10.1016/j.celrep.2022.110519.

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 U S A *114*, 3473-3478. 10.1073/pnas.1617636114.

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. 10.1101/gr.124644.111.

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. 10.1093/bfgp/elab026.

Miotto, B., Ji, Z., and Struhl, K. (2016). Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers. Proc Natl Acad Sci U S A *113*, E4810-4819. 10.1073/pnas.1609060113.

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

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 e977. 10.1016/j.molcel.2018.10.011.

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. 10.1016/j.molcel.2015.03.005.

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. 10.1038/ncomms10208.

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

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. 10.1126/science.1164096.

Prendergast, L., McClurg, U.L., Hristova, R., Berlinguer-Palmini, R., Greener, S., Veitch, K., Hernandez, I., Pasero, P., Rico, D., Higgins, J.M.G., et al. (2020). Resolution of R-loops by INO80 promotes DNA replication and maintains cancer cell proliferation and viability. Nat Commun *11*, 4534. 10.1038/s41467-020-18306-x.

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

Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res *42*, W187-191. 10.1093/nar/gku365.

Shimbo, T., Du, Y., Grimm, S.A., Dhasarathy, A., Mav, D., Shah, R.R., Shi, H.D., and Wade, P.A. (2013). MBD3 Localizes at Promoters, Gene Bodies and Enhancers of Active Genes. Plos Genetics *9*. ARTN e1004028 10.1371/journal.pgen.1004028.

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

Topal, S., Van, C., Xue, Y., Carey, M.F., and Peterson, C.L. (2020). INO80C Remodeler Maintains Genomic Stability by Preventing Promiscuous Transcription at Replication Origins. Cell Rep *32*, 108106. 10.1016/j.celrep.2020.108106.

Venters, B.J., and Pugh, B.F. (2013). Genomic organization of human transcription initiation complexes. Nature *502*, 53-58. 10.1038/nature12535.

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 e2976. 10.1016/j.molcel.2021.05.024.

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. 10.1016/j.cell.2013.03.035.

Xie, L., Pelz, C., Wang, W., Bashar, A., Varlamova, O., Shadle, S., and Impey, S. (2011). KDM5B regulates embryonic stem cell self-renewal and represses cryptic intragenic transcription. EMBO J *30*, 1473-1484. 10.1038/emboj.2011.91.

Xue, Y., Pradhan, S.K., Sun, F., Chronis, C., Tran, N., Su, T., Van, C., Vashisht, A., Wohlschlegel, J., Peterson, C.L., et al. (2017). Mot1, Ino80C, and NC2 Function Coordinately to Regulate Pervasive Transcription in Yeast and Mammals. Mol Cell *67*, 594-607 e594. 10.1016/j.molcel.2017.06.029.

Yamashita, R., Sathira, N.P., Kanai, A., Tanimoto, K., Arauchi, T., Tanaka, Y., Hashimoto, S., Sugano, S., Nakai, K., and Suzuki, Y. (2011). Genome-wide

characterization of transcriptional start sites in humans by integrative transcriptome analysis. Genome Res *21*, 775-789. 10.1101/gr.110254.110.

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. 10.26508/lsa.201800086.

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. 10.1093/nar/gkab1146.

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. 10.1186/gb-2008-9-9-r137.

Chapter 3: INO80 and MOT1 regulate head-on transcription units in Non-Small Cell Lung cancer.

INO80 and MOT1 regulate head-on transcription units in Non-Small Cell Lung cancer.

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

[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

Abstract

Past work has identified the INO80 chromatin-remodeling complex as a critical component of Non-Small Cell Lung cancer (NSCLC) growth. However, it is unclear how INO80 mechanistically supports NSCLC progression. Recent work has demonstrated that INO80, along with the transcriptional regulator MOT1, prevents genotoxic head-on transcription-replication collisions (HO TRCs) in yeast, thus preserving cell viability. Moreover, the recent discovery of temporally regulated head-on transcription units (HO TUs) in tumor cells suggests that transcription is actively regulated in cancer to avoid HO TRCs. We hypothesized that INO80 and MOT1 function to silence HO TUs in NSCLC, thereby preventing HO TRCs. Utilizing genomic assays, we find that INO80 and MOT1 co-bind HO TUs in the A549 NSCLC cell line, and cooperatively silence transcription at a subset of these loci. Furthermore, we find that INO80 and MOT1 cooperate to facilitate NSCLC growth, as well as prevent replication and R-loop dependent DNA damage. This study supports a novel regulatory axis that functions to

silence genotoxic transcription in NSCLC, and thus protect the tumor cell genome from growth-arresting damage.

Introduction

INO80 is an ATP-dependent 15-subunit chromatin remodeling complex that regulates several DNA metabolic processes, including transcription, replication, and repair (Poli et al., 2017). INO80 has recently emerged as a key player in Non-Small Cell Lung cancer (NSCLC) physiology. INO80 complex subunits are genetically amplified at a high frequency in NSCLC histologies and are overexpressed across NSCLC cell lines relative to healthy controls, suggesting that clonal evolution during tumorigenesis selects for increased INO80 activity (Zhang et al., 2017). Clinically, high INO80 expression correlates with worse prognosis in lung cancer patients, demonstrating that INO80 is functionally important for tumor growth (Zhang *et al.*, 2017). In line with this observation, INO80 inhibition markedly reduces the growth of NSCLC cell lines in-vitro, as well as mouse xenografts (Zhang *et al.*, 2017). It is unclear how INO80 mechanistically supports NSCLC growth. Past studies in melanoma and NSCLC models suggested that INO80 binds at enhancers and facilitates chromatin opening, leading to the expression of cancer-associated genes (Zhang *et al.*, 2017; Zhou et al., 2016). Alternatively, work in colorectal, prostate, and breast cancer cells has found that INO80 functions to prevent replication fork stalling and intra-S phase DNA damage, although how it does so remains unclear (Lee et al., 2017; Prendergast et al., 2020; Vassileva et al., 2014). Given the clinical success of genotoxic agents in NSCLC, it is tempting to

speculate that INO80 might be functioning as a DNA protectant in this cancer subtype, and thus serve as an intriguing therapeutic target.

Head-on transcription-replication collisions (HO TRCs) occur when the DNA replisome collides into a converging RNA polymerase during S-phase (Helmrich et al., 2013). When HO TRCs occur over R-loop forming sequences (RLFS), nascent RNA extending outside the RNAPII catalytic core re-hybridizes to the template strand, forming a three-stranded nucleic acid structure known as an R-loop (Hamperl et al., 2017; Helmrich *et al.*, 2013; Zardoni et al., 2021). R-loop formation in turn stabilizes G-quadruplexes on the replisome's leading strand, preventing replication fork progression, and potentially leading to eventual fork collapse into DNA breaks (Hamperl *et al.*, 2017; Kumar et al., 2021; Lee et al., 2020). The genotoxic fallout of HO TRCs leads to cell cycle arrest and senescence in tumor cells, demonstrating the growth-arresting consequence of these events (Nojima et al., 2018). Recent work by our lab has found that RLFS-containing head-on transcription units (HO TUs) form proximal to replication initiation sites in the MCF-7 breast cancer cell line, and are silenced during S-phase, likely to avoid HO TRCs (Kronenberg and Carey, in review). However, the ubiquity of HO TUs across cancer types is unclear. Moreover, the transcriptional regulators of HO TUs are unknown. HO TU regulators could potentially function as essential DNA protectants in fast cycling tumor cells.

Past studies investigating INO80's function as a transcriptional regulator have uncovered a conserved ability to silence non-coding, or pervasive transcription at

replication initiation sites (RIS) (Topal et al., 2020; Xue et al., 2017). Work in mouse embryonic stem cells (mESCs) found that INO80, cooperatively with the TBP antagonist MOT1, silenced pervasive transcription near RIS loci (Topal *et al.*, 2020). However, whether this transcription was head-on in nature, or generated DNA damage upon upregulation, was not assessed. Concomitant work in yeast found that INO80 and MOT1 cooperatively silence head-on transcription at RIS and prevent local replication stress-dependent DNA breaks, demonstrating that INO80 and MOT1 function as DNA protectants through suppressing head-on transcription in this model organism (Topal *et al.*, 2020). However, it is currently unknown whether INO80/MOT1 regulate head-on transcription in human cancers.

In this study, we sought to test the hypothesis that INO80 and MOT1 silence HO TUs in NSCLC, thereby suppressing genotoxic collisions. We find that HO TUs commonly occur on the A549 NSCLC genome, in agreement with observations made in the MCF-7 breast cancer cell line. Utilizing genomic assays, we find that A549 HO TUs are co-bound by INO80 and MOT1, which cooperatively silence a subset with high H2A.Z levels. Similarly, we find that INO80 and MOT1 cooperatively support NSCLC growth and prevent bulk DNA damage, suggesting that INO80 and MOT1 support NSCLC tumorigenesis through preventing HO TRCs. In further support of this model, we find that DNA damage generated by INO80/MOT1 co-depletion is both replication and R-loop dependent. Finally, we find that INO80 depletion generates DNA damage in the A549 NSCLC model, but not healthy lung epithelial BEAS-2B cells, highlighting INO80 as a potentially attractive therapeutic target in NSCLC. In aggregate, this study supports

a model by which INO80 and MOT1 facilitate NSCLC tumorigenesis through preventing HO TRCs.

Results

HO TUs occur on the NSCLC genome

Past work by our lab revealed the presence of head-on transcription units (HU TUs) adjacent to a stringently selected RIS subset in MCF-7 breast cancer cells (Kronenberg and Carey, in review). To decipher whether A549 NSCLC cells also harbor HO TUs within their genome, we utilized a similar workflow as was done in the MCF-7 analysis (Supplemental Figure 1A). Briefly, we first identified a high confidence RIS subset through intersecting ~700bp loci that have demonstrated conserved replication initiation activity across cell types ('core origins'), A549 EdU-seq peaks, which represent A549-specific replication initiation hotspots (Macheret and Halazonetis, 2019), and loci containing an A549-specific epigenetic signature predictive of origin-of-replication complex binding (Miotto et al., 2016), yielding a final set of 5,277 RIS. Called RIS were positioned at replication timing profile inverted V-apexes (Supplemental Figure 1B, left panel), were enriched for early replicating regions (Supplemental Figure 1C, right panel), and showed expected positional profiles (Supplemental Figure 1C,D). The presence of HO TUs at this RIS subset was then assessed. To do this, we utilized A549 CAGE-seq data, which maps transcription start sites genome wide through cap isolation and sequencing (Yan et al., 2022). We defined HO TUs as regions bookended on one end by a head-on A549 CAGE-seq peak within 1kb of an RIS border, and on the other the RIS summit (Figure 1A). We found that 3,271 of the 5,277 RIS contained at least

one HO TU (Figure 1B), in agreement with observations in MCF-7 cells. In total, we identified 3,886 HO TUs, due to the formation of multiple units at some RIS. To investigate HO TU transcriptional activity, we utilized A549 PRO-seq data, which maps directional transcription via nuclear run-on methodology (Mahat et al., 2016). Viewing CAGE-seq and PRO-seq signals at HO TUs on browser tracks clearly demonstrates the presence of units of head-on transcription at individual RIS (Figure 1C). Viewing CAGE-seq and PRO-seq signals across all HO TUs on a heatmap clearly demonstrates that head-on transcription is initiating at and elongating within the TUs (Figure 1D). Thus, in agreement with past observations made in MCF-7 cells, head-on transcription is a feature of a majority of A549 RIS, and occurs within distinct, identifiable units. It is unclear why ~38% of the RIS subset do not contain an HO TU. Given that transcription at RIS has been found to correlate with earlier replication timing, it is possible that these loci harbor RIS that fire relatively later in S-phase. Alternatively, these loci could be false positives due to the inherent difficulties in mapping RIS locations (Ganier et al., 2019).

Past work in MCF-7 cells demonstrated that HO TUs were mostly pervasive in nature (Kronenberg and Carey, in review). We next assessed whether A549 HO TUs were associated with pervasive transcripts. 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 PRO-seq data as has been done previously (Jacquier, 2009; Liu et al., 2022) (Figure 1E). We then categorized HO TUs by whether they overlapped with any of these pervasive transcript classes. We found that 7% of HO TU associations

were with PROMPTs, 23% with eRNAs, 13% with asTSSa, 44% with sTSSa, and 13% with transcripts outside these classes (Figure 1F). Although we can't distinguish whether sTSSa-associated transcripts are pervasive in nature or are part of 5' gene transcription due to the co-directional relationship between these two transcript types, over 50% of HO TU associations are with bonafide pervasive transcripts, demonstrating that a majority of A549 HO TUs are pervasive in nature, in agreement with the observations made in MCF-7 cells (Kronenberg and Carey, in review).

HO TRCs are especially toxic when they occur over R-loop forming sequences (RLFS), which enable nascent RNA re-hybridization to the template strand and highly stable fork stalling (Hamperl *et al.*, 2017; Kumar *et al.*, 2021). To address whether A549 HO TUs contain RLFS within their template strand, we utilized a dataset containing bioinformatically predicted RLFS from R-loopDB (Jenjaroenpun et al., 2017). We found that 3,058 of the 3,886 total HO TUs contained a RLFS within their template strand (Supplemental Figure 2A). When viewing RLFS density across HO TUs on a heatmap, we found that RLFS localized within the TU, peaking in the TU center (Supplemental Figure 2B). Finally, we evaluated RLFS density at HO TUs across transcript class associations. We found that RLFS densities were similar across subsets, with sTSSa-associated HO TUs showing relatively increased RLFS density, in agreement with their positioning near RLFS-rich gene TSS (Chen et al., 2017). Notably, RIS that did not contain HO TUs were devoid of RLFS (data not shown), suggesting that either RLFS are a prerequisite for transcription, or this RIS subset is largely false positives, as CG-rich DNA is a typical feature of RIS loci (Akerman et al., 2022). Collectively, this analysis

demonstrates that RLFS are a feature of A549 HO TUs, and suggests that HO TRCs stemming from HO TU transcription would likely be genotoxic in nature (Supplemental Figure 2D).

INO80 and MOT1 cooperatively silence a subset of HO TUs

Based on previous work done in yeast and mESCs, we hypothesized that INO80 and MOT1 bind at HO TUs and function to silence transcription in A549 cells (Xue at al., 2017; Topal et al., 2020). To test this hypothesis, we first evaluated INO80 and MOT1 occupancy at HO TU loci. To do this, we performed chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq), using antibodies against the ACTR5 subunit of the INO80 complex and MOT1 (Tosi et al., 2013). Browser track examples show clear local events of INO80 and MOT1 co-binding at HO TUs, biased to the TSS side (Figure 2A). Heatmap visualization of ChIP-seq signals across HO TUs demonstrates that INO80 and MOT1 bind at the TSS on a global scale (Figure 2B). Thus, INO80 and MOT1 co-bind at HO TU TSSs in A549 cells.

We next evaluated INO80 and MOT1's transcriptional regulatory activity at A549 HO TUs. To do this, we depleted INO80 and MOT1 over a 72 hour period via siRNA transfection (Supplemental Figure 3A), and then extracted cells for nascent RNA-seq , which measures chromatin-associated RNA levels genome-wide (Bhatt et al., 2012). As a control, we used cells transfected with a scramble siRNA over the same time period. Local browser track examples of INO80 and MOT1 co-bound HO TUs clearly show that INO80 and MOT1 co-depletion increases transcription at individual loci (Figure 2C). To

assess transcriptional changes at HO TUs globally, we mapped the log2 fold-change in nascent RNA-seq reads per 50bp bin on all HO TUs and performed k-means clustering using python deeptools. We found that 2 well-defined clusters formed (Figure 2D). The first cluster (C1) contains HO TUs that on average exhibited ~2-fold upregulation upon INO80 and MOT1 co-depletion. The second cluster (C2) contains HO TUs that appear largely unaffected by co-depletion. To systematically assess transcriptional effects across clusters, we evaluated the HO TU RPKM distributions from control and IM-depleted cells. We found that IM-depletion significantly shifted RPKMs up at C1 HO TUs but did not significantly change C2 HO TU transcription (Supplemental Figure 3B). Likewise, differential expression analysis demonstrated that C1 HO TU transcriptional changes are significantly skewed towards upregulation, whereas C2 HO TU changes show little bias towards upregulation (Supplemental Figure 3C). Collectively, this analysis demonstrates that INO80 and MOT1 function either additively, cooperatively, or in isolation to silence a subset of HO TUs. To parse this out, we performed nascent RNA-seq in single knockdown backgrounds (Supplemental Figure 3A) and evaluated transcriptional change at C1 HO TUs. In agreement with observations made in yeast, we found that while INO80 single knockdown caused increased HO TU transcription, MOT1 single knockdown generated no effect. However, INO80 and MOT1 co-depletion generates a synergistic increase in transcription, demonstrating that INO80 and MOT1 function cooperatively to silence A549 HO TUs (Supplemental Figure 3D), in agreement with studies done in yeast (Topal *et al.*, 2020; Xue *et al.*, 2017).

We next sought to understand what molecular features define the subset of INO80/MOT1-regulated HO TUs. Interestingly, we found that INO80 and MOT1 occupancy levels were similar at C1 and C2 HO TUs, suggesting INO80/MOT1 binding is not sufficient for HO TU silencing (Supplemental Figure 4A). We reasoned that INO80/MOT1 regulatory activity might be due to an interaction with a particular chromatin feature. INO80 binds the histone variant H2A.Z with high affinity and promotes histone exchange with H2A to stimulate transcriptional silencing (Papamichos-Chronakis et al., 2011). Interestingly, INO80 removal of H2A.Z is critical for the maintenance of genome integrity in yeast (Papamichos-Chronakis *et al.*, 2011). We thus reasoned that INO80/MOT1 regulated HO TUs might be enriched for H2A.Z relative to the non-regulated subset. To assess this, we utilized a publicly available A549 H2A.Z ChIP-seq dataset to evaluate H2A.Z enrichment at C1 and C2 HO TUs. This analysis revealed that C1 HO TUs contain significantly higher levels of H2A.Z relative to C2 HO TUs (Supplemental Figure 4A). Collectively, this analysis suggests that HO TU regulation is not determined by INO80 and MOT1 binding per se, but instead the interplay of INO80 and MOT1 with the local chromatin environment, specifically H2A.Z-containing nucleosomes.

Finally, we evaluated whether INO80/MOT1-regulated HO TUs were enriched in a particular pervasive transcript class. When comparing association frequencies between C1 and C2 HO TUs, we found that PROMPTs were overrepresented in C1, while all other classes were either similar or underrepresented (Supplemental Figure 4B). To systematically assess INO80/MOT1 regulatory activity across classes, we first

compared subset HO TU RPKM distributions between control and co-depleted conditions (Supplemental Figure 4C). Interestingly, we found that on a population level, only PROMPT-associated HO TUs showed a significant upregulation in INO80/MOT1 depleted conditions. In agreement with earlier analysis, this PROMPT-specific regulatory activity was independent of INO80/MOT1 occupancy (Supplemental Figure 4D). Differential expression analysis of HO TU transcription by subset similarly demonstrates that only PROMPT-associated HO TUs experience highly skewed upregulation upon INO80/MOT1 depletion (Supplemental Figure 4E). In aggregate, this analysis demonstrates that INO80 and MOT1 uniformly silence PROMPT-associated HO TUs. INO80/MOT1's variable activity at HO TUs associated with other transcript classes suggests that additional factors are required at these units that do not occur uniformly across the transcript population.

INO80 and MOT1 cooperatively prevent replication and R-loop dependent DNA damage in NSCLC

Upregulation of HO TU transcription via INO80/MOT1 inhibition could potentially increase the frequency of HO TRCs. Given the genotoxic nature of HO TRCs, we next evaluated whether INO80/MOT1 inhibition could generate DNA damage in A549 cells. We depleted INO80 and MOT1 individually or together over 72 hours (about 3 cell cycles), harvested whole cell extracts, and performed a western blot against the DNA damage marker ɣH2Ax. We found that while individual depletion of INO80 significantly upregulated ɣH2Ax, individual depletion of MOT1 had no effect on ɣH2Ax levels. However, co-depletion of INO80/MOT1 generated a synergistic increase in ɣH2Ax

levels, demonstrating that INO80 and MOT1 cooperatively prevent DNA damage in A549 cells (Figure 3A). Interestingly, these DNA damage phenotypes across single and double knockdown backgrounds mirrored the effects of single and double knockdowns on HO TU upregulation, as well as cell growth (Figure 3B). Collectively, these analyses suggest that INO80 and MOT1 cooperatively prevent DNA damage through their HO TU silencing activity, and this activity supports tumor cell growth.

If INO80 and MOT1 were preventing DNA damage via mitigating HO TRCs, then damage induced by INO80/MOT1 co-depletion should be replication-dependent in nature. To assess this, we utilized Palbociclib, a CDK4/6 inhibitor, to induce G1 arrest in A549 cells. After 24 hours of Palbociclib treatment, ~98% of A549 cells were arrested in G1-phase, demonstrating the feasibility of this approach (Figure 3C). We next co-depleted INO80/MOT1 for 72 hours, and then treated cells with either DMSO or Palbociclib for 24 hours prior to performing a whole cell extraction and western blot for ɣH2Ax. We found that Palbociclib treatment partially rescued DNA damage induced by INO80/MOT1 co-depletion, demonstrating that INO80 and MOT1 prevent damage in a replication-dependent manner (Figure 1D).

Our previous analysis found that A549 HO TUs are enriched in RLFS in the template strand, suggesting HO TU dysregulation could result in HO TRCs over RLFS (Supplemental Figure 2). Past work has demonstrated that an RLFS at head-on collision sites is necessary for induction of DNA damage, as collisions stabilize R-loops at these sites, leading to formation of secondary structures in the leading strand and

stable replication fork stalling (Hamperl *et al.*, 2017; Kumar *et al.*, 2021). Moreover, overexpression of RnaseH, an R-loop nuclease, has been shown to ameliorate DNA damage generated by induced collisions (Hamperl *et al.*, 2017). We reasoned that if INO80/MOT1 were preventing HO TRCs, then DNA damage caused by co-depletion should be rescued by overexpression of RNaseH. To test this, we co-depleted INO80/MOT1 for 72 hours, and then performed either a mock transfection or transfected a RnaseH overexpression plasmid for 24 hours prior to performing a whole cell extraction and western blot for ɣH2Ax. We found that RNaseH transfection starkly rescued DNA damage induced by INO80/MOT1 co-depletion, demonstrating that INO80 and MOT1 prevent DNA damage in an R-loop dependent manner (Figure 3E). Collectively, these analyses suggest that INO80 and MOT1 function as DNA protectants in the A549 cell line through preventing HO TRCs (Figure 3F).

INO80 is a NSCLC-specific DNA protectant

Past work has found that INO80 depletion selectively inhibits NSCLC tumor cell line growth, with little effect on healthy lung cells (Zhang *et al.*, 2017). However, the molecular basis for this selective activity remains unclear. Given the enhanced vulnerability of tumor cells to exogenously introduced replication stress, we postulated that INO80 might be functioning as a NSCLC-specific DNA protectant (Dobbelstein and Sorensen, 2015). To test this hypothesis, we depleted INO80 in both NSCLC A549 cells and immortalized lung epithelial BEAS-2B cells, and separately evaluated both ɣH2Ax and p53 induction. Importantly, our siRNA method achieved similar depletion efficiencies across cell lines (Figure 4A) and recapitulated the growth phenotypes seen

previously (Figure 4B). We found that INO80 depletion exclusively upregulated ɣH2Ax and p53 in A549 cells, demonstrating that INO80 functions as an NSCLC-specific DNA protectant (Figure 4C). As an orthogonal approach, we evaluated A549 and BEAS-2B specific gene expression changes upon INO80 depletion across a set of consensus p53-activated genes (Fischer, 2017). We found that INO80 depletion in A549 cells significantly upregulated 11 genes in the set, while only 2 genes were upregulated in BEAS-2B cells, further reinforcing the NSCLC-selective nature of INO80's function as a DNA protectant (Figure 4D). Similarly, gene set enrichment analysis (GSEA) across apoptosis and p53 related gene sets demonstrated that INO80 depletion significantly upregulates these pathways in A549, but not BEAS-2B cells (Figure 1E) (Subramanian et al., 2005). Finally, we hypothesized that INO80's selective activity might stem from differences in basal replication stress levels between cell lines. To investigate this, we quantified replication stress through evaluating the expression of a transcriptional signature that correlates with oncogene-induced replication stress load (Guerrero Llobet et al., 2022). We found that this signature was more highly expressed in INO80 relative to BEAS-2B cells, suggesting that INO80 is interacting with cancer-specific endogenous replication stress to function as a DNA protectant (Figure 4F).

Discussion

HO TRCs are potent genotoxic events that lead to the loss of cycling cell viability (Hamperl *et al.*, 2017; Nojima *et al.*, 2018). The recent discovery of temporally regulated HO TUs in breast cancer cells suggested that transcriptional regulators might function in cancer to prevent head-on collisions during S-phase (Kronenberg and Carey, in review).

However, whether HO TUs occur in other cancer types, or what these regulators are remains unknown. Our analysis, utilizing an integrated bioinformatic and wet-lab strategy, reveals that HO TUs occur at a high frequency in the A549 NSCLC model cell line, suggesting that HO TUs are a general feature of tumor cells. Moreover, we find that the transcriptional regulators INO80 and MOT1 bind at A549 HO TU loci and cooperatively silence transcription at a subset marked by high H2A.Z, an INO80 substrate (Papamichos-Chronakis *et al.*, 2011). Furthermore, INO80 and MOT1 cooperatively prevent replication and R-loop dependent DNA damage, suggesting functional prevention of HO TRCs. INO80 is a multi-functional protein complex that has been shown to directly remove RNAPII as well as R-loops from chromatin (Lafon et al., 2015; Prendergast *et al.*, 2020). While we cannot rule out a post-collision mechanism that explains damage prevention, the participation of MOT1, which has no known functions outside transcriptional regulation (Auble et al., 1994), strongly suggests that the prevention of HO TU transcription plays a role in suppressing HO TRC-induced damage. It is possible that combined transcriptional and post-transcriptional activity by INO80 is necessary to fully prevent HO TRC-induced damage, thus affecting the afferent and efferent arms of a collision.

The direct result of a HO TRC is the generation of a stalled replication fork (Kumar *et al.*, 2021). Stalled forks become uncoupled from the MCM helicase, leading to the generation of single-stranded DNA tracts (ssDNA) (Saxena and Zou, 2022; Toledo et al., 2017). ssDNA formation stimulates the binding of the heterotrimeric protein Replication Protein A (RPA), which in turn recruits effector molecules such as ATR and

ATM to stabilize and resolve the fork (Toledo *et al.*, 2017). This RPA catalyzed signaling cascade also leads to intra-S phase checkpoint activation, leading to suppression of origin firing, and conservation of the RPA pool (Toledo *et al.*, 2017). Under conditions of extreme replication stress, RPA becomes exhausted, leading to the presence of unstable forks and pan-DNA breakage, an event known as replication catastrophe (RC) (Toledo *et al.*, 2017). If INO80 prevents fork stalling, then it might only function as a DNA protectant in conditions of elevated replication stress, where the RPA pool has little capacity to buffer an increase in stalled forks. Given the rapid mutation rate and elevated stress levels observed in NSCLC , we reasoned that INO80 might prevent DNA damage in a cancer-specific manner. Indeed, past work found that INO80 depletion was selectively toxic to NSCLC cell lines, but not healthy lung epithelial cells (Zhang *et al.*, 2017). We find that INO80 prevents the formation of the DNA damage markers ɣH2Ax and p53 in the A549 NSCLC cell line, but not the healthy lung epithelial BEAS-2B cell line, supporting the idea that INO80 is a cancer-specific DNA protectant. Interestingly, A549 cells show an increase in replication stress over BEAS-2B cells, as ascertained by a transcriptional signature (Guerrero Llobet *et al.*, 2022). Collectively, these findings suggest that inhibition of the INO80/MOT1 regulatory axis could induce tumor-selective genotoxicity through interacting with the cancer-specific replication stress hallmark.

Acknowledgements

This work was supported by NIH grants R01 GM074701 to M.F.C., and the TRDRP 2019B Predoctoral fellowship award T30DT0906 to M.K.

Figures

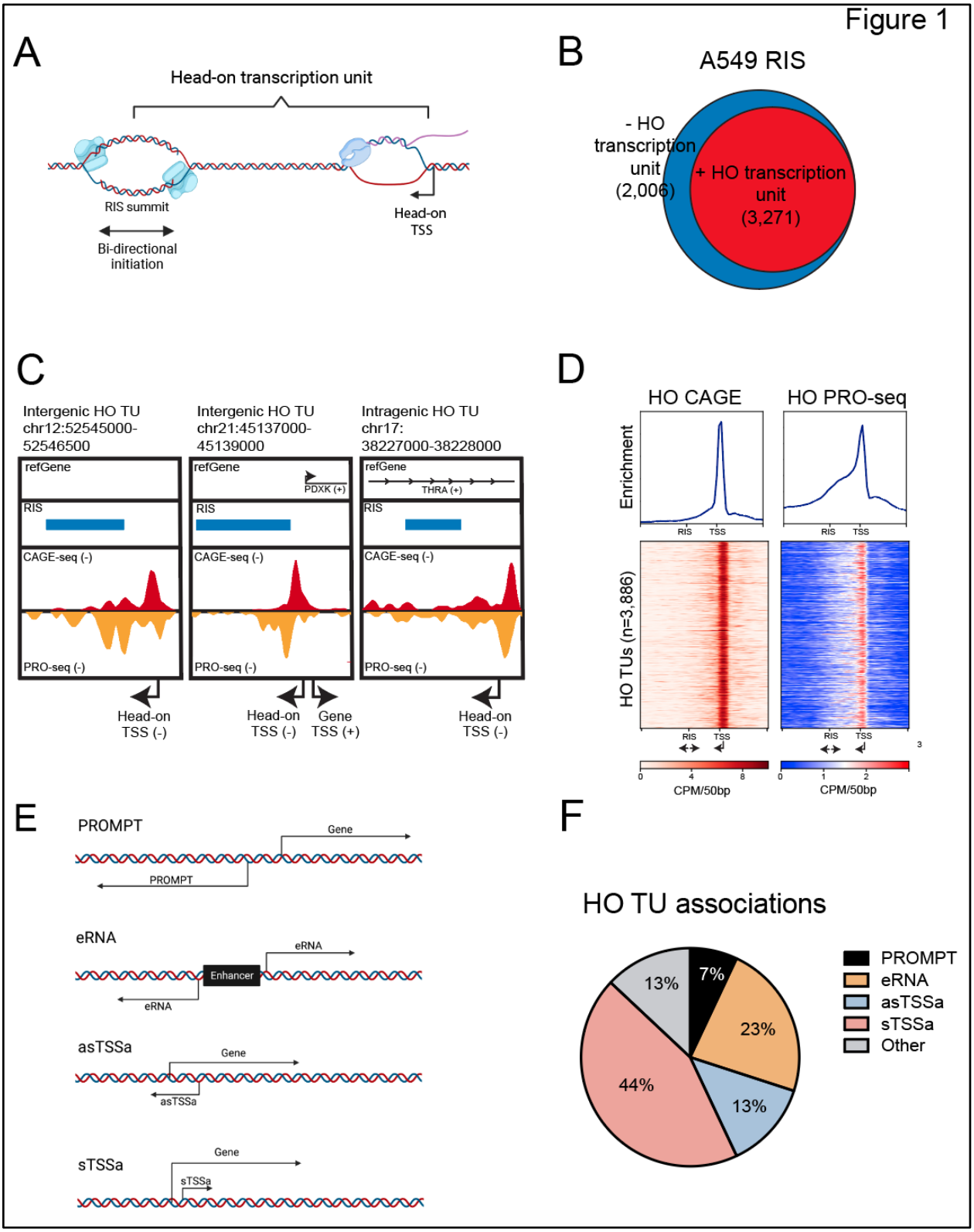


Figure 1
A
Head-on transcription unit
RIS summit
Bi-directional initiation
Head-on TSS
B
A549 RIS
- HO transcription unit (2,006)
+ HO transcription unit (3,271)
C
Intergenic HO TU chr12:52545000-52546500
Intergenic HO TU chr21:45137000-45139000
Intragenic HO TU chr17: 38227000-38228000
refGene
RIS
CAGE-seq (-)
PRO-seq (-)
PDXK (+)
THRA (+)
Head-on TSS (-)
Gene TSS (+)
D
HO CAGE
HO PRO-seq
Enrichment
HO TUs (n=3,886)
RIS
TSS
CPM/50bp
E
PROMPT
Gene
eRNA
Enhancer
asTSSa
sTSSa
F
HO TU associations
7%
23%
13%
44%
13%
PROMPT
eRNA
asTSSa
sTSSa
Other

Figure 2

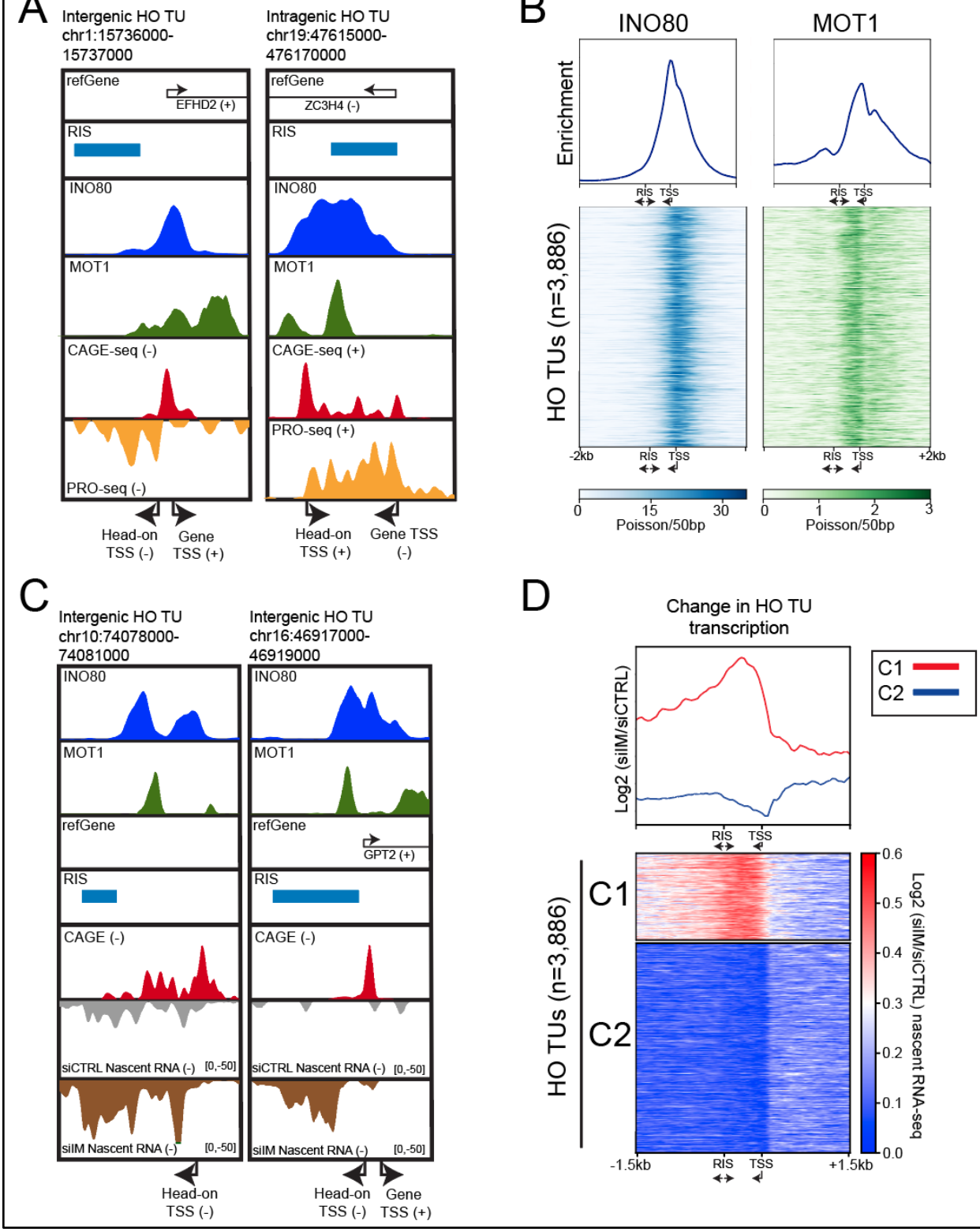

A
Intergenic HO TU
chr1:15736000-
15737000
Intragenic HO TU
chr19:47615000-
476170000
refGene
EFHD2 (+)
ZC3H4 (-)
RIS
INO80
MOT1
CAGE-seq (-)
CAGE-seq (+)
PRO-seq (-)
PRO-seq (+)
Head-on
TSS (-)
Gene
TSS (+)
Head-on
TSS (+)
Gene TSS
(-)
B
INO80
MOT1
Enrichment
HO TUs (n=3,886)
RIS
TSS
-2kb
+2kb
0
15
30
Poisson/50bp
0
1
2
3
Poisson/50bp
C
Intergenic HO TU
chr10:74078000-
74081000
Intergenic HO TU
chr16:46917000-
46919000
GPT2 (+)
CAGE (-)
siCTRL Nascent RNA (-)
[0,-50]
siIM Nascent RNA (-)
[0,-50]
Head-on
TSS (-)
Head-on
TSS (-)
Gene
TSS (+)
D
Change in HO TU
transcription
Log2 (siIM/siCTRL)
C1
C2
HO TUs (n=3,886)
-1.5kb
+1.5kb
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Log2 (siIM/siCTRL) nascent RNA-seq

Figure 3

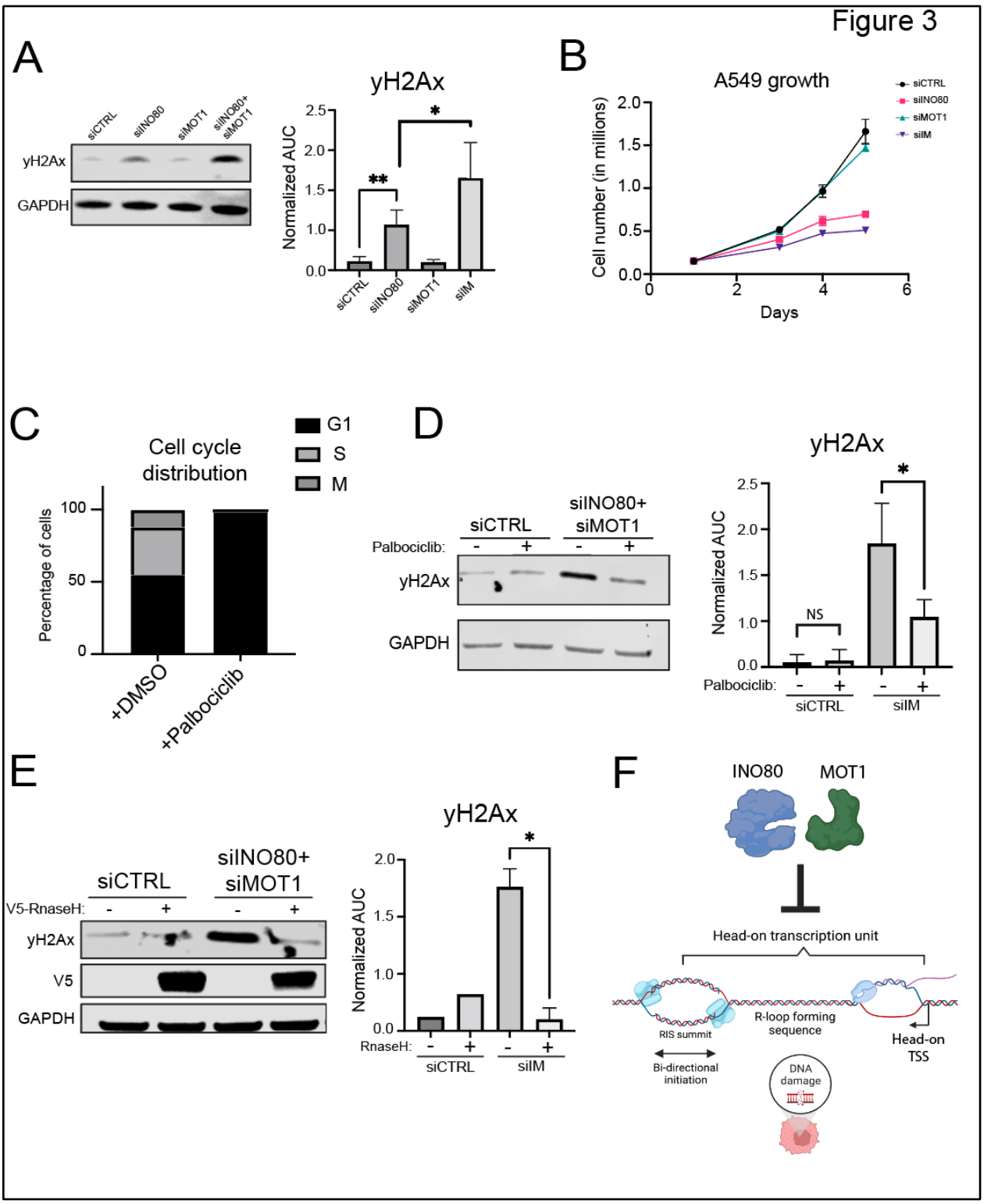

A
siCTRL
siINO80
siMOT1
siINO80+ siMOT1
γH2Ax
GAPDH
γH2Ax
Normalized AUC
siCTRL
siINO80
siMOT1
siIM
B
A549 growth
siCTRL
siINO80
siMOT1
siIM
Cell number (in millions)
Days
C
Cell cycle distribution
G1
S
M
Percentage of cells
+DMSO
+Palbociclib
D
siCTRL
siINO80+ siMOT1
Palbociclib:
γH2Ax
GAPDH
γH2Ax
Normalized AUC
NS
Palbociclib:
siCTRL
siIM
E
siCTRL
siINO80+ siMOT1
V5-RnaseH:
γH2Ax
V5
GAPDH
γH2Ax
Normalized AUC
RnaseH:
siCTRL
siIM
F
INO80
MOT1
Head-on transcription unit
RIS summit
R-loop forming sequence
Head-on TSS
Bi-directional initiation
DNA damage

Figure 4

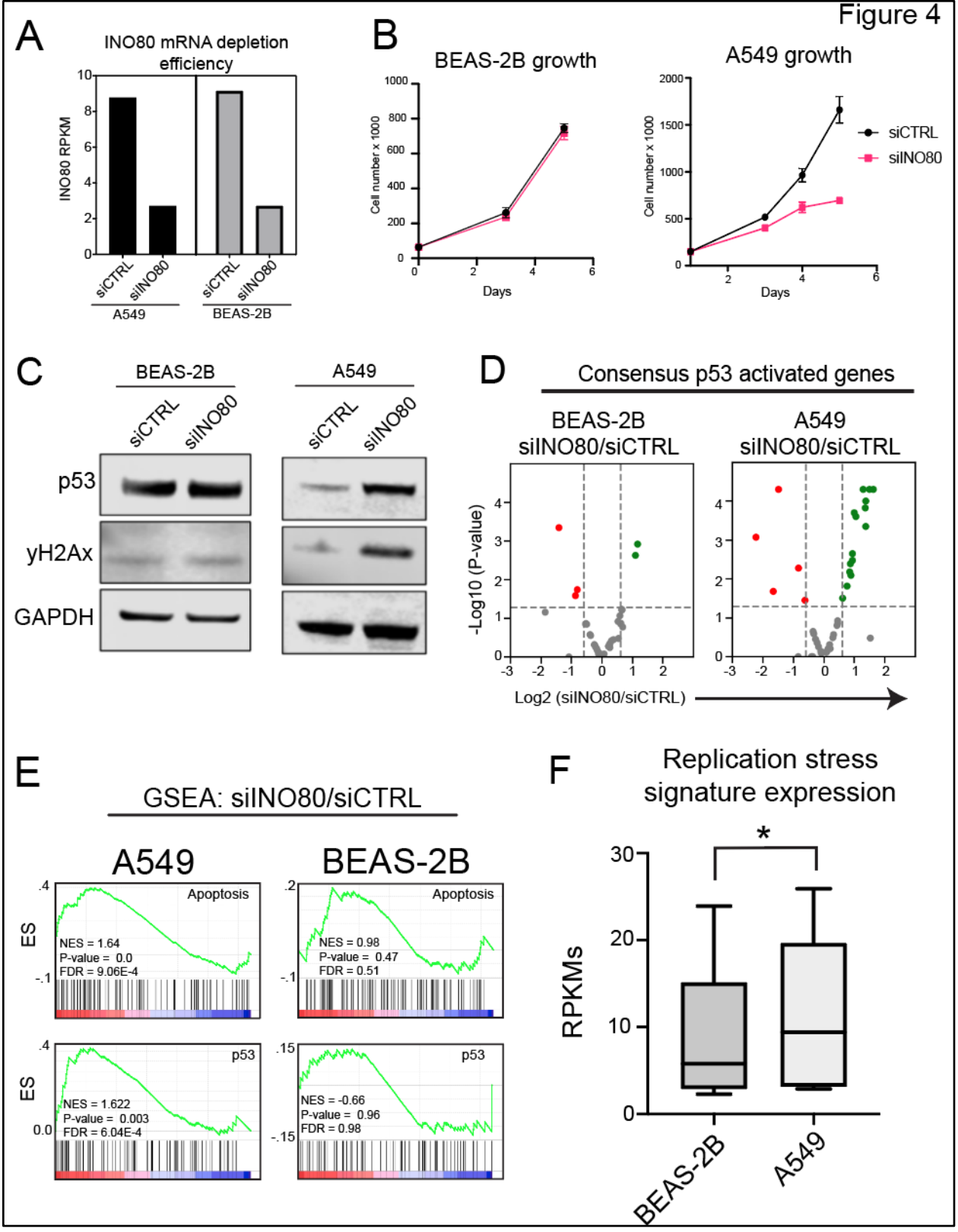

A
INO80 mRNA depletion efficiency
INO80 RPKM
siCTRL
siINO80
A549
BEAS-2B
B
BEAS-2B growth
A549 growth
Cell number x 1000
Days
siCTRL
siINO80
C
BEAS-2B
A549
siCTRL
siINO80
p53
γH2Ax
GAPDH
D
Consensus p53 activated genes
BEAS-2B siINO80/siCTRL
A549 siINO80/siCTRL
-Log10 (P-value)
Log2 (siINO80/siCTRL)
E
GSEA: siINO80/siCTRL
A549
BEAS-2B
Apoptosis
NES = 1.64
P-value = 0.0
FDR = 9.06E-4
NES = 0.98
P-value = 0.47
FDR = 0.51
p53
NES = 1.622
P-value = 0.003
FDR = 6.04E-4
NES = -0.66
P-value = 0.96
FDR = 0.98
ES
F
Replication stress signature expression
RPKMs
BEAS-2B
A549

Figure Legends

Figure 1: HO TUs occur on the NSCLC genome

A. Graphic representation of a head-on transcription unit (HO TU). B. Diagram of total A549 RIS demarcated by the presence or absence of at least one HO TU. C. Browser track examples of A549 HO TUs. D. Average profiles and heatmaps of A549 head-on CAGE-seq and PRO-seq enrichment at distance normalized HO TU loci. E. Graphic representation of pervasive transcript classes. F. Pie chart showing the percentage of total A549 HO TU associations with a given pervasive transcript class.

Figure 2: INO80 and MOT1 cooperatively silence a subset of HO TUs

A. Browser track examples of INO80 and MOT1 co-binding at A549 HO TU TSSs. B. Average profiles and heatmaps of A549 INO80 and MOT1 ChIP-seq enrichment at distance normalized HO TUs. C. Browser track examples of A549 HO TUs that are silenced by INO80 and MOT1. D. Average profile and heatmap of the log2 fold change in head-on nascent RNA-seq signal between INO80/MOT1 co-depleted and control conditions at distance normalized A549 HO TUs subset by k-means clustering.

Figure 3: INO80 and MOT1 cooperatively prevent replication and R-loop dependent DNA damage in NSCLC.

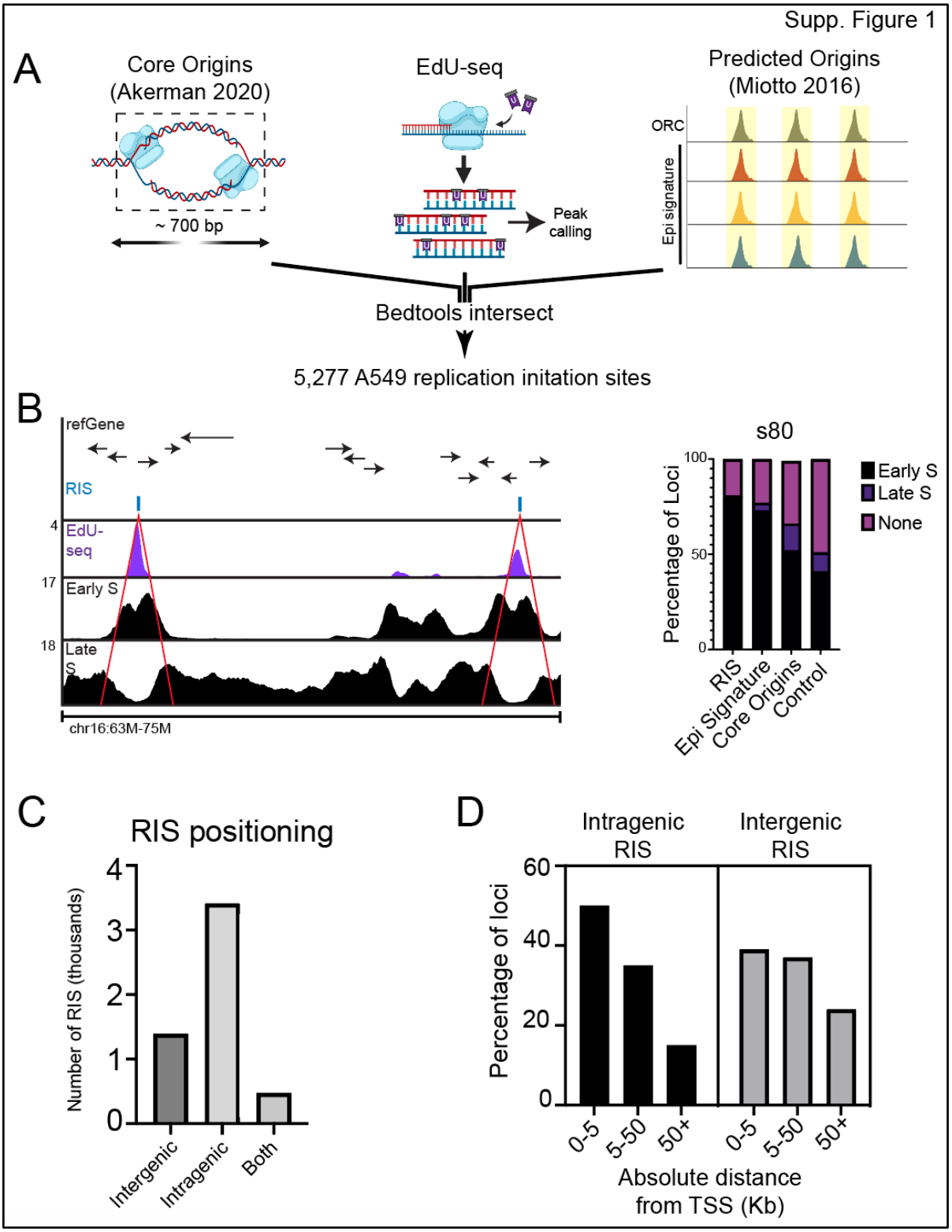
A. Western blot quantifying bulk ɣH2Ax levels across single and double knockdown conditions in A549 cells (Left). Bar graph showing normalized ɣH2Ax signal across western blot replicates (Right) (AUC: Area under the curve as assessed by ImageJ). B. Growth curve of A549 cells in single and double knockdown conditions. C. Bar graph of

A549 cell cycle distribution after 24 hours of either DMSO or Palbociclib treatment. D. Western blot quantifying bulk ɣH2Ax levels across replication competent and replication arrested conditions in A549 cells (Left). Bar graph showing normalized ɣH2Ax signal across western blot replicates (Right) (AUC: Area under the curve as assessed by ImageJ). E. Western blot quantifying bulk ɣH2Ax levels across mock and RNaseH overexpressing conditions in A549 cells (Left). Bar graph showing normalized ɣH2Ax signal across western blot replicates (Right) (AUC: Area under the curve as assessed by ImageJ). F. Graphic representation of modeled INO80 and MOT1 function at HO TUs.

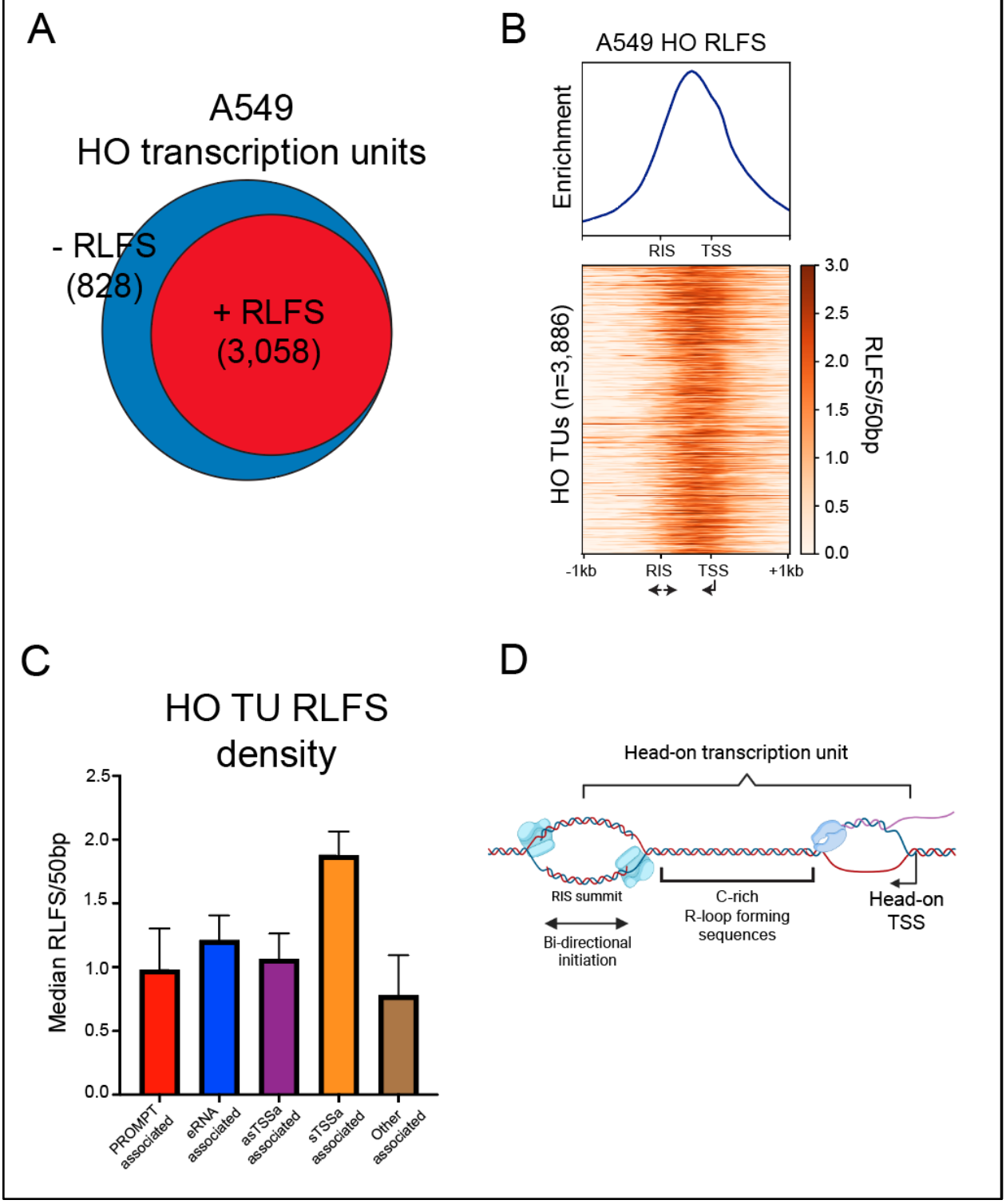
Figure 4: INO80 is a NSCLC-specific DNA protectant

A. Bar graph showing the depletion efficiency of INO80 mRNA in both A549 and BEAS-2B cells 72 hours after INO80 siRNA transfection. B. Growth curve of A549 (Left) and BEAS-2B (Right) cells in control and INO80 depleted conditions. C. Western blot quantifying bulk ɣH2Ax and p53 levels in control and INO80 depleted conditions in BEAS-2B (Left) and A549 (Right) cells. D. Volcano plots showing the differential expression of a p53-activated gene set in BEAS-2B (Left) and A549 (Right) cells. E. Gene set enrichment analysis of INO80 depletion-induced changes in gene expression for apoptosis and p53-related genes as annotated by the PANTHER classification system across A549 cells (Left) and BEAS-2B cells (Right). F. Box and whisker plot of the RPKM distribution of a replication stress transcriptional signature in BEAS-2B and A549 cells.

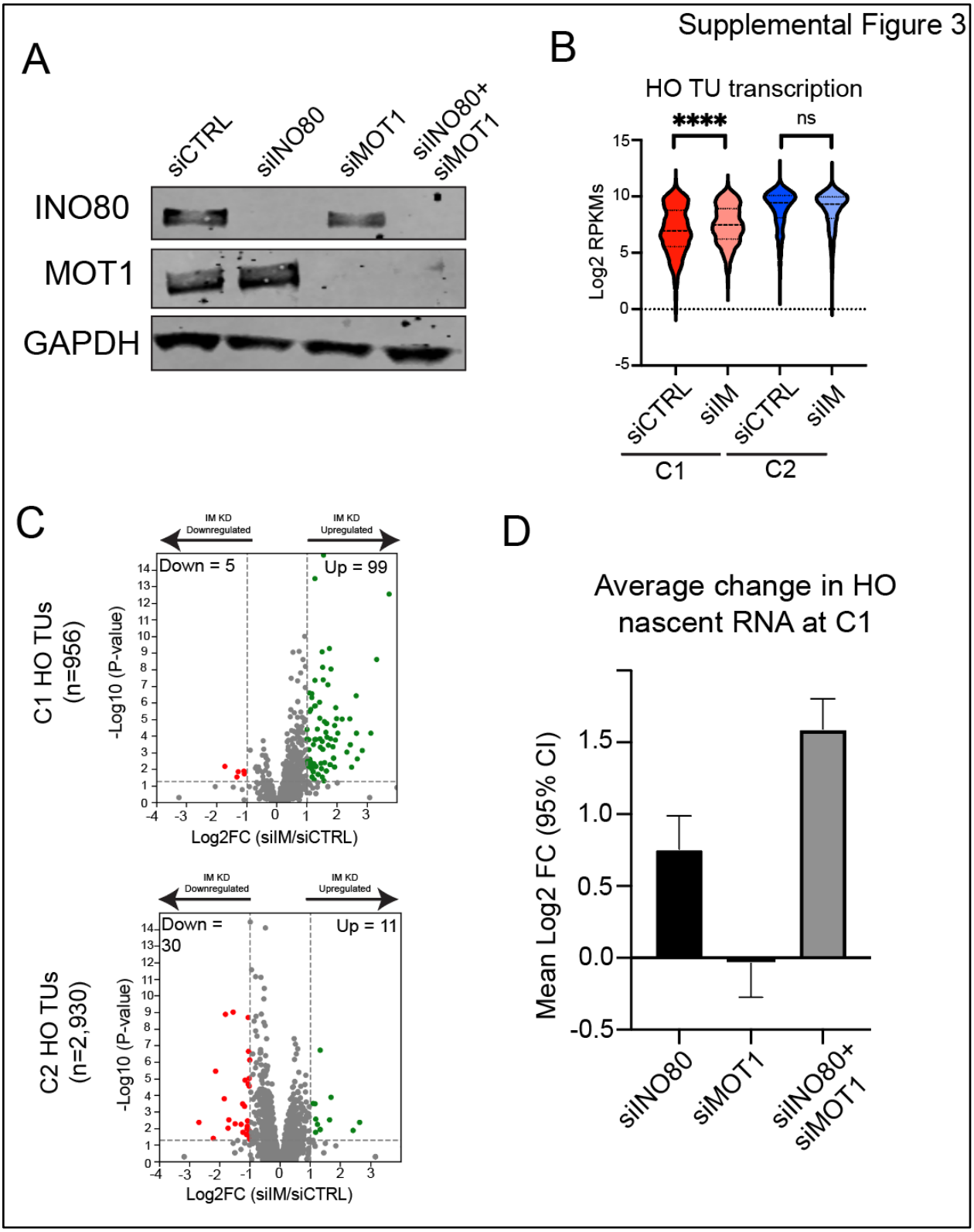
Supplemental Figures



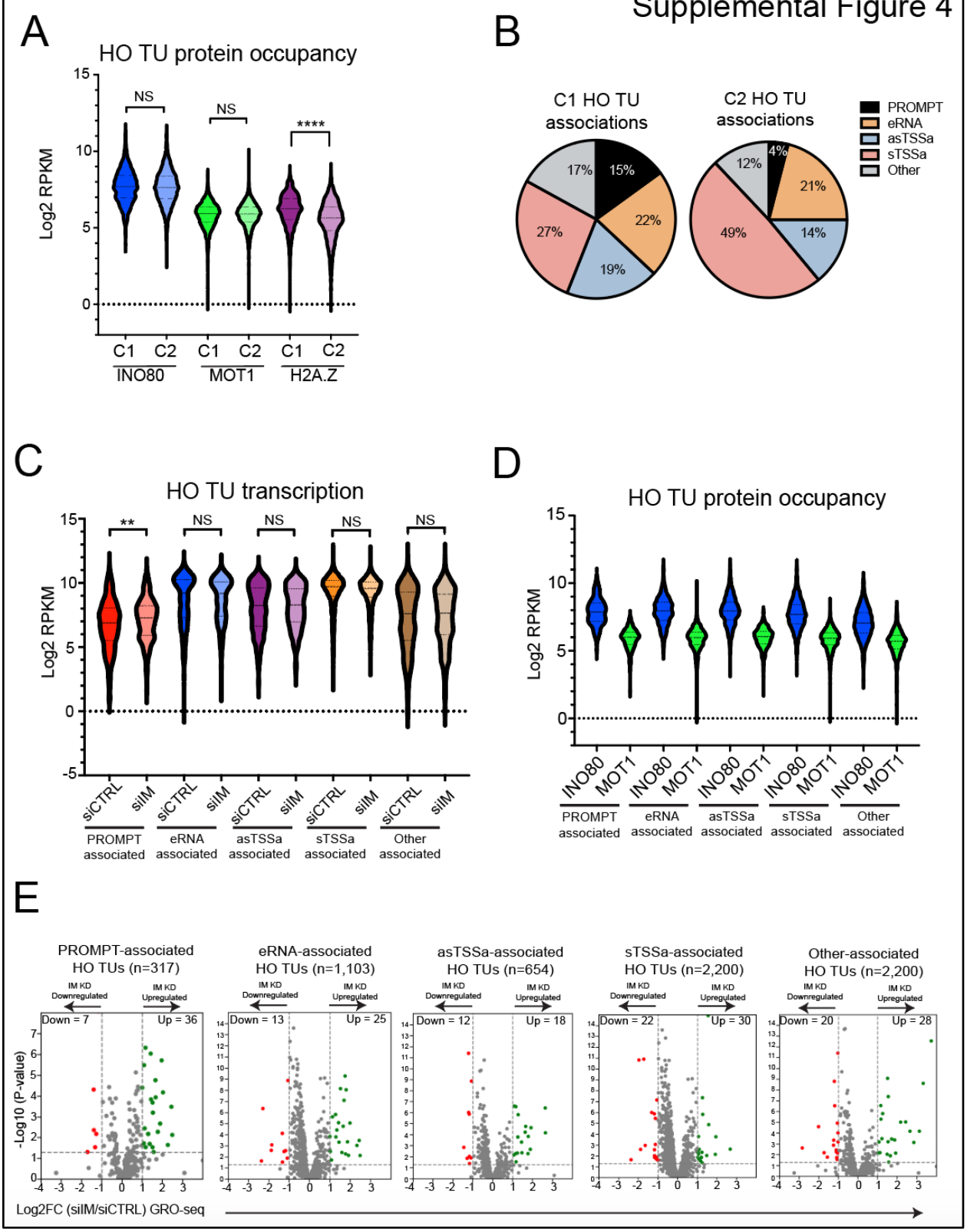
Supplemental Figure 2



Supplemental Figure 3



Supplemental Figure 4



Supplemental Figure Legends

Supplemental Figure 1: Identification of high confidence A549 RIS.

A. Schematic of the strategy used to identify A549 RIS. B. (Left) Browser track showing RIS (top track, blue markers), EdU-seq enrichment (second track from top) and replication timing profiles (Bottom 2 tracks). Red lines demarcate inverted-V structures. (Right) Distribution of s80 labels across RIS, benchmark, and control datasets. C. Bar graph showing RIS frequency by position relative to gene bodies. D. Bar graphs showing RIS frequency by absolute distance relative to the nearest protein-coding TSS.

Supplemental Figure 2: A549 HO TUs are enriched in R-loop forming sequences.

A. Diagram of total HO TUs demarcated by the presence or absence of at least one RLFS in the template strand. B. Average profile and heatmap of RLFS frequency on template strand within 50 bp bins at distance normalized HO TUs. C Bar chart showing the median RLFS density across HO TUs subset by pervasive transcript class association. D. Graphic representation of HO TUs with positioned R-loop forming sequences within the transcribed body.

Supplemental Figure 3: INO80 and MOT1 cooperatively silence a subset of HO TUs.

A. Western blot of INO80 and MOT1 protein levels 72 hours after single siRNA depletion or co-depletion in A549 cells. B. Violin plots of nascent RNA-seq RPKM distribution at clustered HO TUs in control and co-depleted conditions. C. Volcano plots showing differential expression analysis of clustered HO TUs. D. Bar chart showing the

mean log2 fold-change (+ 95% C.I.) in nascent RNA-seq signal at C1 HO TUs between single and double knockdown conditions.

Supplemental Figure 4: Features of INO80/MOT1 regulated HO TUs.

A. Violin plots of INO80, MOT1, and H2A.Z ChIP-seq RPKM distributions at clustered HO TUs. B. Pie charts showing the percentage of total C1 (Left) or C2 (Right) A549 HO TU associations with a given pervasive transcript class . C. Violin plots of nascent RNA-seq RPKM distributions in control and co-depleted conditions at HO TUs subset by pervasive transcript association. D. Violin plots of INO80 and MOT1 ChIP-seq RPKM distributions as HO TUs subset by pervasive transcript association. E. Volcano plots showing differential expression analysis of HO TUs subset by pervasive transcript association.

Materials and methods

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
ACTR5/Arp5	Proteintech	Cat# 21505-1-AP, RRID:1234
INO80	Proteintech	Cat# 18810-1-AP, RRID: NA
MOT1	Abcam Abcam	Cat# ab196491, RRID: NA Cat# ab72285, RRID NA
ɣH2Ax	Millipore sigma Abcam	Cat# 05-636, RRID: NA Cat# ab81299, RRID: NA
P53	Cell Signaling	Cat# 18032S, RRID: NA
GAPDH	Santa Cruz Biotechnology	Cat# sc-365062, RRID:NA
V5	Abcam	Cat# ab15828, RRID:NA
Plasmids		
ppyCAG_RnaseH1_WT	Addgene	Cat# 111906
siRNA		
INO80	Horizon Discovery	Cat# L-004176-01-0005
MOT1	Horizon Discovery	Cat# L-012628-00-0005
Small molecules		

Palbociclib	Selleck Chemicals	Cat# S1116
Biotin-Azide	Thermo FIsher	Cat# B10184
EdU	Thermo Fisher	Cat# A10044
Kits		
KAPA RNA Hyper+RiboErase HMR	Roche diagnostics	Cat# 08098131702
KAPA mRNA Hyper Prep	Roche diagnostics	Cat# 08098115702
KAPA HyperPrep Kit	Roche diagnostics	Cat# 07962312001
Click-iT™ EdU Alexa Fluor™ 647 Flow Cytometry Assay	Thermo Fisher	Cat# C10424
Experimental Models: Cell Lines		
A549	ATCC	CCL-185
BEAS-2B	ATCC	CRL-9609
Deposited Data		
Core origin coordinate file	Akerman et al. 2020	NCBI Gene Expression Omnibus (GEO): GSE128477
A549 Repli-seq	David Gilbert, FSU	ENCODE: doi:10.17989/ENCSR594FTB

A549 CAGE-seq	Yan et al. 2022	NCBI Gene Expression Omnibus (GEO): GSE132660
A549 PRO-seq	John Lis, Cornell	ENCODE: doi:10.17989/ENCSR244HDV
Software and Algorithm		
Bedtools		Quinlan and Hall 2010
Samtools		Li et al., 2009
Tophat2		Kim et al., 2013
MACS2		Zhang et al., 2008
Bowtie2		Langmead et al., 2009
Deeptools		Ramirez et al., 2014
HOMER		Heinz et al., 2010

EXPERIMENTAL PROCEDURES

EdU-seq

Asynchronous A549 cells were pulsed with 10uM EdU for 30 minutes, followed by cell harvesting and processing using the Click-iT™ EdU Alexa Fluor™ 647 Flow Cytometry Assay. Biotin Azide was substituted for Alexa Fluor, and post-click reaction the samples were processed as was done in (Macheret and Halazonetis, 2019). Sequencing libraries were made using the KAPA HyperPrep Kit. Fastq files were mapped to the hg19 genome with bowtie2 (Langmead et al., 2009) to generate bam alignment files. All bam files were then processed with samtools (Li et al., 2009) so that duplicates were

removed, and low-quality reads were filtered out. MACS2 peakcall (Zhang et al., 2008) was then used to call peaks from EdU-seq samples relative to an input. Peaks were merged using bedtools merge (Quinlan and Hall, 2010) if summits were within 5000 bp of each other.

RIS identification

Core origin summits (Akerman *et al.*, 2020) A549 EdU-seq peaks, and Dnase-seq peaks from loci containing overlapping peaks of A549 H3K27ac ChIP-seq, A549 H3K4me2 ChIP-seq, and A549 Dnase-seq were extended 1 kb in each direction using bedtools slop (Consortium, 2012; Quinlan and Hall, 2010). These extended peaks were then intersected using bedtools intersect (Quinlan and Hall, 2010). Intersected core origin coordinates were used to represent RIS.

RIS validation

Samtools bedcov (Li *et al.*, 2009) was used to map reads from A549 replication timing datasets (Repli-seq) (Consortium, 2012) to RIS regions and comparator dataset loci (Epigenetic signature peaks, core origins, and randomly selected Dnase-seq peaks). For epigenetic signature and Dnase-seq peaks, the center of each peak was extended 1kb in each direction for mapping using bedtools slop (Quinlan and Hall, 2010). For RIS and core origins, the center of all coordinate locations were taken and extended 1kb in each direction for mapping using bedtools slop. 5,277 random Dnase-seq peaks were selected through using bedtools shuffle (Quinlan and Hall, 2010) 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 80% of the total reads from one fraction, then it was marked with an s80 label for that fraction as was done previously (Dellino et al., 2013).

RIS sub-setting by intragenic or intergenic status

Intragenic RIS were identified by using bedtools intersect to find RIS entirely confined within protein-coding gene body termini as annotated from the GENCODE database (Frankish et al., 2019). Intergenic RIS were identified by using bedtools subtract (Quinlan and Hall, 2010) to identify the remaining RIS. If RIS both overlapped gene body regions and adjacent intergenic regions, they were categorized as 'both'.

RIS sub-setting by TSS distance

The HOMER annotatePeaks function (Heinz et al., 2010) was used to determine the distance from the nearest protein-coding TSS for each RIS location based off the RIS center coordinate. RIS were then binned by the calculated absolute distance.

Head-on transcription unit (HO TU) identification

Directional NET CAGE-seq peaks (Hirabayashi et al., 2019) were intersected with regions delimited by a RIS center and 1kb downstream of the RIS border proximal to the NDR using bedtools intersect. Minus strand NET CAGE-seq peaks were intersected with RIS that formed a downstream NDR, and plus strand NET CAGE-seq peaks were intersected with RIS that formed an upstream NDR. Intersected peaks were labeled HO

TU TSS, and the cognate RIS center point represented the HO TU terminus. Some RIS contained multiple HO TUs due to multiple NET CAGE-seq peaks intersecting with the demarcated RIS region.

CAGE-seq data processing

Fastq files were mapped to the hg19 genome with bowtie2 (Langmead *et al.*, 2009) to generate bam alignment files. All bam files were then processed with samtools (Li *et al.*, 2009) so that duplicates were removed, and low-quality reads were filtered out. Replicate bam files were merged for downstream analysis using samtools merge (Li *et al.*, 2009). Merged and QC'd bam files were separated by strand and MACS2 peakcall (Zhang *et al.*, 2008) was then used to identify stranded peaks and generate a bedgraph file. This bedgraph file was then converted to bigwig files describing mapped reads in counts per million using the bedGraphtoBigWig script from ENCODE (Consortium, 2012; Kent et al., 2010) for downstream analysis using the python deeptools software suite (Ramirez et al., 2014).

PRO-seq data processing

Raw fastq files from were mapped to the hg19 genome with tophat2 (Kim et al., 2013) to produce bam alignment files. Duplicates and low quality reads were removed from bam files via samtools (Li *et al.*, 2009). Replicate bam files were merged for downstream analysis using samtools merge (Li *et al.*, 2009). 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 (Ramirez *et al.*, 2014).

Pervasive transcript identification

A549 PRO-seq datasets were used to perform de novo transcript discovery via HOMER software, yielding 29,518 transcripts. Transcripts were labeled as PROMPTs if they were intergenic, within 5kb upstream of a TSS, and were antisense to the proximal gene. This yielded 2,108 total PROMPTs. Transcripts were labeled as eRNAs if their TSS overlapped with enhancer regions called by the ROSE software with gene TSS exclusion (Whyte et al., 2013). This yielded 8,779 total eRNAs. Transcripts were labeled as asTSSa if they overlapped with TSS plus 500bp downstream and were divergent to gene direction. This yielded 4,317 asTSSa. Transcripts were labeled as sTSSa if they overlapped with TSS plus 500bp downstream and were in the same direction as gene transcription. This yielded 10,840 transcriptsTSSa.

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.

RLFS identification and association with features

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

(Jenjaroenpun *et al.*, 2017). The merged RLFS coordinate file for the hg19 genome was downloaded and separated by strand. Concomitantly, RIS were subset by accessible region location based on the plus strand as was done in prior analyses. To identify RIS that contained RLFS in the head-on transcription template strand, bedtools intersect was used to find subset RIS that overlapped with the directionally appropriate stranded RLFS file. Resulting files were then concatenated. The same pipeline was used to assess RLFS presence within the template strand of HO TUs.

To determine RLFS high and low RIS, a bedgraph file describing RLFS frequency per 50bp 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 along with RIS coordinates for python deeptools analysis. Output files describing RLFS density within individual RIS units were rank-ordered and the bottom 25% and top 25% loci were selected for low and high groups respectively.

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

ChIP-seq

ChIP-seq was performed in A549 cells as was done in (Xue *et al.*, 2017), using an antibody against the INO80C subunit ACTR5 and MOT1. Sequencing libraries were

made using the KAPA HyperPrep Kit. Fastq files were mapped to the hg19 genome with bowtie2 (Langmead *et al.*, 2009) to generate bam alignment files. All bam files were then processed with samtools (Li *et al.*, 2009) so that duplicates were removed, and low-quality reads were filtered out. MACS2 peakcall (Zhang *et al.*, 2008) was then used to generate read normalized treatment and background bedgraph files from IP and input controls respectively. MACS2 bdgcmp (Zhang *et al.*, 2008) 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 (Consortium, 2012; Kent et al., 2010) for downstream analysis using the python deeptools software suite (Ramirez et al., 2014).

siRNA transfection

A549 or BEAS-2B cells were reverse transfected at 150,000 cells/well with 0.5 ug of the appropriate siRNA using 6ul lipofectamine siRNA max and antibiotic free media. Transfection media was exchanged the next day. Cells were incubated for 72 hours prior to harvesting for downstream analysis.

Nascent RNA-seq

Nascent RNA-seq was performed in A549 cells as was done in Bhatt et al., 2017. Sequencing libraries were made with the KAPA RNA Hyper+Riboerase kit. Raw fastq files were mapped to the hg19 genome with tophat2 (Kim *et al.*, 2013) to produce bam alignment files. Duplicates and low quality reads were removed from bam files via

samtools (Li *et al.*, 2009). Replicate bam files were merged for downstream analysis using samtools merge (Li *et al.*, 2009). 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 (Ramirez *et al.*, 2014). Bam files were also converted to stranded bam files for downstream analysis.

Immunoblotting

Cells were harvested and washed twice with cold dPBS. Cells were then lysed in RIPA buffer containing protease and phosphatase inhibitors per standard protocol, and extracts were mixed with 1x SDS loading buffer. Extracts were size separated on an SDS-PAGE gel via electrophoresis, and transferred with the iblot2 transfer system. Blots were blocked with LiCOR blocking buffer and incubated overnight at four degrees with the appropriate diluted antibodies. Blots were washed and incubated with secondary LiCOR antibodies for band imaging on a LiCOR machine. Blots were quantified by imageJ with an area under the curve (AUC) score for the protein of interest and loading control, and replicate AUC ratios were combined for statistical analysis.

Growth Curves

Cells were reverse transfected at 150,000 cells per well with the appropriate siRNA as described above. Measurements were taken across triplicate samples at 24 hours, 72 hours, and 120 hours post-transfection using a TC-10 cell counter.

Induction of cell cycle arrest

A549 cells were treated with Palbociclib dissolved in DMSO at a concentration of 2uM. Cells were harvested after a 24 hour incubation period for downstream analysis. Propidium Iodide staining was performed to validate cell cycle arrest in the G1-phase.

Plasmid transfection

One ug of the ppyCAG_RnaseH1_WT plasmid was forward transfected per well of a 6-well plate containing A549 cells using 2ul lipofectamine 2000 and antibiotic free media. A549 cells were transfected at 80% confluency. Transfection media was exchanged after 4 hours. Cells were harvested after a 24 hour incubation period for downstream analysis.

mRNA-seq

Total RNA from A549 or BEAS-2B cells was extracted via trizol and purified via Dnase treatment and a second round of trizol extraction. Purified RNA was then used as input for processing with the KAPA mRNA HyperPrep kit. Raw fastq files were mapped to the hg19 genome with tophat2 (Kim *et al.*, 2013) to produce bam alignment files. Duplicates and low quality reads were removed from bam files via samtools (Li *et al.*, 2009). Replicate bam files were merged for downstream analysis using samtools merge (Li *et al.*, 2009). Merged and QC'd bam files were then used as inputs for cufflinks to generate RPKM files at annotated genes.

Differential expression analysis

Tag directories from replicate bam files of interest 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 .05 and a log2 fold change of 1 or -1.

Gene Set Enrichment Analysis

Gene set enrichment analysis was performed as described in (Subramanian *et al.*, 2005).

RPKM calculations

Merged and QC'd bam files generated from fastq files as previously described were converted to sam files, separated by strand, reconverted to bam files, and indexed using samtools (Li *et al.*, 2009). To find RPKMs, samtools bedcov was used to map reads from stranded bam files directionally onto desired regions. Mapped reads within 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. All violin plot RPKM visualizations were generated via PRISM 9 statistical software.

Cell culture

A549 and BEAS-2B cells were cultured on TC qualified plates in media containing DMEM/F12 (1:1) (Thermo Fisher, 11320-033), supplemented with 10% fetal bovine serum.

Graphics generation

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

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.

References

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. 10.1038/s41467-020-18527-0.

Auble, D.T., Hansen, K.E., Mueller, C.G., Lane, W.S., Thorner, J., and Hahn, S. (1994). Mot1, a global repressor of RNA polymerase II transcription, inhibits TBP binding to DNA by an ATP-dependent mechanism. Genes Dev *8*, 1920-1934. 10.1101/gad.8.16.1920.

Bhatt, D.M., Pandya-Jones, A., Tong, A.J., Barozzi, I., Lissner, M.M., Natoli, G., Black, D.L., and Smale, S.T. (2012). Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. Cell *150*, 279-290. 10.1016/j.cell.2012.05.043.

Chen, L., Chen, J.Y., Zhang, X., Gu, Y., Xiao, R., Shao, C., Tang, P., Qian, H., Luo, D., Li, H., et al. (2017). R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters. Mol Cell *68*, 745-757 e745. 10.1016/j.molcel.2017.10.008.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74. 10.1038/nature11247.

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. 10.1101/gr.142331.112.

Dobbelstein, M., and Sorensen, C.S. (2015). Exploiting replicative stress to treat cancer. Nat Rev Drug Discov *14*, 405-423. 10.1038/nrd4553.

Fischer, M. (2017). Census and evaluation of p53 target genes. Oncogene *36*, 3943-3956. 10.1038/onc.2016.502.

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. 10.1101/gr.183848.114.

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. 10.1093/nar/gky955.

Guerrero Llobet, S., Bhattacharya, A., Everts, M., Kok, K., van der Vegt, B., Fehrmann, R.S.N., and van Vugt, M. (2022). An mRNA expression-based signature for oncogene-induced replication-stress. Oncogene *41*, 1216-1224. 10.1038/s41388-021-02162-0.

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 e719. 10.1016/j.cell.2017.07.043.

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. 10.1016/j.molcel.2010.05.004.

Helmrich, A., Ballarino, M., Nudler, E., and Tora, L. (2013). Transcription-replication encounters, consequences and genomic instability. Nat Struct Mol Biol *20*, 412-418. 10.1038/nsmb.2543.

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. 10.1038/s41588-019-0485-9.

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

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. 10.1093/nar/gkw1054.

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. 10.1093/bioinformatics/btq351.

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. 10.1186/gb-2013-14-4-r36.

Kumar, C., Batra, S., Griffith, J.D., and Remus, D. (2021). The interplay of RNA:DNA hybrid structure and G-quadruplexes determines the outcome of R-loop-replisome collisions. Elife *10*. 10.7554/eLife.72286.

Lafon, A., Taranum, S., Pietrocola, F., Dingli, F., Loew, D., Brahma, S., Bartholomew, B., and Papamichos-Chronakis, M. (2015). INO80 Chromatin Remodeler Facilitates

Release of RNA Polymerase II from Chromatin for Ubiquitin-Mediated Proteasomal Degradation. Mol Cell *60*, 784-796. 10.1016/j.molcel.2015.10.028.

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. 10.1186/gb-2009-10-3-r25.

Lee, C.Y., McNerney, C., Ma, K., Zhao, W., Wang, A., and Myong, S. (2020). R-loop induced G-quadruplex in non-template promotes transcription by successive R-loop formation. Nat Commun *11*, 3392. 10.1038/s41467-020-17176-7.

Lee, S.A., Lee, H.S., Hur, S.K., Kang, S.W., Oh, G.T., Lee, D., and Kwon, J. (2017). INO80 haploinsufficiency inhibits colon cancer tumorigenesis via replication stress-induced apoptosis. Oncotarget *8*, 115041-115053. 10.18632/oncotarget.22984.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079. 10.1093/bioinformatics/btp352.

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. 10.1016/j.celrep.2022.110519.

Macheret, M., and Halazonetis, T.D. (2019). Monitoring early S-phase origin firing and replication fork movement by sequencing nascent DNA from synchronized cells. Nat Protoc *14*, 51-67. 10.1038/s41596-018-0081-y.

Mahat, D.B., Kwak, H., Booth, G.T., Jonkers, I.H., Danko, C.G., Patel, R.K., Waters, C.T., Munson, K., Core, L.J., and Lis, J.T. (2016). Base-pair-resolution genome-wide

mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc *11*, 1455-1476. 10.1038/nprot.2016.086.

Miotto, B., Ji, Z., and Struhl, K. (2016). Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers. Proc Natl Acad Sci U S A *113*, E4810-4819. 10.1073/pnas.1609060113.

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 e977. 10.1016/j.molcel.2018.10.011.

Papamichos-Chronakis, M., Watanabe, S., Rando, O.J., and Peterson, C.L. (2011). Global regulation of H2A.Z localization by the INO80 chromatin-remodeling enzyme is essential for genome integrity. Cell *144*, 200-213. 10.1016/j.cell.2010.12.021.

Poli, J., Gasser, S.M., and Papamichos-Chronakis, M. (2017). The INO80 remodeller in transcription, replication and repair. Philos Trans R Soc Lond B Biol Sci *372*. 10.1098/rstb.2016.0290.

Prendergast, L., McClurg, U.L., Hristova, R., Berlinguer-Palmini, R., Greener, S., Veitch, K., Hernandez, I., Pasero, P., Rico, D., Higgins, J.M.G., et al. (2020). Resolution of R-loops by INO80 promotes DNA replication and maintains cancer cell proliferation and viability. Nat Commun *11*, 4534. 10.1038/s41467-020-18306-x.

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

Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res *42*, W187-191. 10.1093/nar/gku365.

Saxena, S., and Zou, L. (2022). Hallmarks of DNA replication stress. Mol Cell *82*, 2298-2314. 10.1016/j.molcel.2022.05.004.

St Germain, C.P., Zhao, H., Sinha, V., Sanz, L.A., Chedin, F., and Barlow, J.H. (2022). Genomic patterns of transcription-replication interactions in mouse primary B cells. Nucleic Acids Res *50*, 2051-2073. 10.1093/nar/gkac035.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550. 10.1073/pnas.0506580102.

Toledo, L., Neelsen, K.J., and Lukas, J. (2017). Replication Catastrophe: When a Checkpoint Fails because of Exhaustion. Mol Cell *66*, 735-749. 10.1016/j.molcel.2017.05.001.

Topal, S., Van, C., Xue, Y., Carey, M.F., and Peterson, C.L. (2020). INO80C Remodeler Maintains Genomic Stability by Preventing Promiscuous Transcription at Replication Origins. Cell Rep *32*, 108106. 10.1016/j.celrep.2020.108106.

Tosi, A., Haas, C., Herzog, F., Gilmozzi, A., Berninghausen, O., Ungewickell, C., Gerhold, C.B., Lakomek, K., Aebersold, R., Beckmann, R., and Hopfner, K.P. (2013). Structure and subunit topology of the INO80 chromatin remodeler and its nucleosome complex. Cell *154*, 1207-1219. 10.1016/j.cell.2013.08.016.

Tubbs, A., Sridharan, S., van Wietmarschen, N., Maman, Y., Callen, E., Stanlie, A., Wu, W., Wu, X., Day, A., Wong, N., et al. (2018). Dual Roles of Poly(dA:dT) Tracts in Replication Initiation and Fork Collapse. Cell *174*, 1127-1142 e1119. 10.1016/j.cell.2018.07.011.

Vassileva, I., Yanakieva, I., Peycheva, M., Gospodinov, A., and Anachkova, B. (2014). The mammalian INO80 chromatin remodeling complex is required for replication stress recovery. Nucleic Acids Res *42*, 9074-9086. 10.1093/nar/gku605.

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. 10.1016/j.cell.2013.03.035.

Xue, Y., Pradhan, S.K., Sun, F., Chronis, C., Tran, N., Su, T., Van, C., Vashisht, A., Wohlschlegel, J., Peterson, C.L., et al. (2017). Mot1, Ino80C, and NC2 Function Coordinately to Regulate Pervasive Transcription in Yeast and Mammals. Mol Cell *67*, 594-607 e594. 10.1016/j.molcel.2017.06.029.

Yan, B., Tzertzinis, G., Schildkraut, I., and Ettwiller, L. (2022). Comprehensive determination of transcription start sites derived from all RNA polymerases using ReCappable-seq. Genome Res *32*, 162-174. 10.1101/gr.275784.121.

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. 10.1093/nar/gkab1146.

Zhang, S., Zhou, B., Wang, L., Li, P., Bennett, B.D., Snyder, R., Garantziotis, S., Fargo, D.C., Cox, A.D., Chen, L., and Hu, G. (2017). INO80 is required for oncogenic transcription and tumor growth in non-small cell lung cancer. Oncogene *36*, 1430-1439. 10.1038/onc.2016.311.

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. 10.1186/gb-2008-9-9-r137.

Zhou, B., Wang, L., Zhang, S., Bennett, B.D., He, F., Zhang, Y., Xiong, C., Han, L., Diao, L., Li, P., et al. (2016). INO80 governs superenhancer-mediated oncogenic transcription and tumor growth in melanoma. Genes Dev *30*, 1440-1453. 10.1101/gad.277178.115.

Chapter 4: Mapping cisplatin-induced DNA damage in Non-Small Cell Lung cancer.

Mapping cisplatin-induced DNA damage in Non-Small Cell Lung cancer.

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

[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

Abstract

The chemotherapeutic cisplatin serves as a backbone in many first-line treatment regimens for Non-Small Cell Lung cancer (NSCLC). A key effort in NSCLC drug development is the discovery of therapies that can safely amplify cisplatin's efficacy. One way this could be rationally achieved is through targeting chromatin-based mechanisms that either limit or increase cisplatin's ability to damage tumor cell DNA. However, the relationship between chromatin states and cisplatin's genotoxic potential in NSCLC is unclear. Here, we generate a high resolution and quantitative genome-wide map of cisplatin-induced DNA-damage in an NSCLC model through performing ChIP-seq on the damage marker ɣH2Ax. Furthermore, we utilize this dataset to assess the relationship between cisplatin-induced DNA damage and chromatin state. This work ultimately provides a resource that can be leveraged to better understand the molecular mechanisms governing cisplatin genotoxicity in NSCLC.

Introduction

Cisplatin is a widely used chemotherapeutic for the treatment of Non-Small Cell Lung cancer (NSCLC) (Fennell et al., 2016). Cisplatin induces anti-tumor toxicity through interacting with DNA and inducing the formation of intrastrand or interstrand crosslinks (ICLs) between GG dinucleotides (Hu et al., 2016; Shu et al., 2016). These crosslinks can either be repaired by the nucleotide excision repair pathway (NER), or generate replication fork stalling during genome replication, in both cases leading to ssDNA formation and activation of the DNA damage response (Duan et al., 2020; Fennell *et al.*, 2016; Frankenberg-Schwager et al., 2005). A key effort in modern NSCLC therapeutic development is finding rational combination partners that can amplify cisplatin toxicity in tumors. Understanding where and at what frequency cisplatin induces DNA damage on the NSCLC genome could reveal chromatin states or epigenetic factors that functionally amplify or suppress cisplatin genotoxicity.

Past work mapping cisplatin integration and nucleotide excision repair (NER) activities genome-wide in lymphocyte cells revealed principles of cisplatin activity on the genome (Hu *et al.*, 2016). Chromatin state analysis in this study found that while cisplatin induced GG intrastrand crosslinks indiscriminately, NER activity mainly occurred at promoter and enhancer regions (Hu *et al.*, 2016). However, it remains unclear how these dynamics result in downstream DNA damage. For example, does cisplatin integrated in regions with poor NER kinetics generate more DNA damage than cisplatin in regions with highly active NER? Furthermore, this study was not powered to assess

interstrand crosslinks induced by cisplatin, which are highly toxic in nature (Hashimoto et al., 2016) . Thus, it remains unclear where and at what frequency cisplatin functionally damages the human genome.

In this study, we map DNA damage induced by cisplatin treatment genome-wide in the NSCLC A549 cell line. To do so, we leverage ɣH2Ax, which marks both double-stranded and single-stranded breaks (DSBs and SSBs) (Marti et al., 2006; Podhorecka et al., 2010). We find that cisplatin induces widespread, focal sites of DNA damage across the NSCLC genome. We further find that damage levels are much higher at both active and bivalent promoters relative to other chromatin states, suggesting that chromatin-based processes at promoter regions contribute to cisplatin genotoxicity. Thus, this work reveals novel relationships between cisplatin activity and the chromatin environment in NSCLC.

Results

Genome-wide mapping of cisplatin-induced damage across the NSCLC genome.

In order to map DNA damage induced by cisplatin at high resolution on the NSCLC genome, we utilized a ChIP-seq based strategy (Figure 1A). Briefly, we treated A549 cells with saline or 100uM cisplatin for 24 hours, followed by cell extraction, crosslinking, and immunoprecipitation of the DNA damage marker ɣH2Ax. Precipitated DNA was then purified, sequenced, normalized to an input, and processed for peak calling. Viewing, ɣH2Ax signal on a browser track demonstrates that cisplatin treatment generates clear and widespread ɣH2Ax peaks (Figure 1B). Global evaluation of the

ɣH2Ax signal at cisplatin-induced peaks shows a clear upregulation upon drug treatment (Figure 1C). Peak calling by MACS2 revealed that cisplatin treatment resulted in 81,342 ɣH2Ax peaks as opposed to 349 peaks in the saline treated samples (Figure 1D). Finally, we evaluated cisplatin peak association with different genomic features. We found that peaks distributed across TSS, gene bodies, and intergenic regions, demonstrating that induced damage is not limited to a single genomic feature (Figure 1E). It is important to note that ɣH2Ax marks both SSBs that occur during NER, as well as DSBs that occur during replication fork collapse (Marti et al., 2006; Podhorecka et al., 2010). Thus, our dataset does not distinguish between these features.

Cisplatin-induced damage across chromHMM identified chromatin states.

To understand how cisplatin genotoxicity interacts with the chromatin environment, we looked at cisplatin-induced ɣH2Ax signals across 18 chromatin states identified by chromHMM in A549 cells (Ernst and Kellis, 2017). Interestingly, we found that damage induction was much higher at both active and poised TSS relative to any other state (Figure 2A). Interestingly, although active TSS had higher levels of transcription (Figure 2B), poised TSS had higher ɣH2Ax levels. This demonstrates that promoters, partially independent of transcriptional activity, contain features that support cisplatin's ability to damage DNA. No other chromatin states displayed clear and uniform enrichment in cisplatin-induced DNA damage, although across states there are clearly loci that host high levels of damage. Thus, it is likely that still unknown determinants of cisplatin genotoxicity, independent of chromatin state, exist on the genome.

Discussion

Since its approval decades ago, cisplatin is still used as a first line treatment for advanced Non-Small Cell Lung cancer (NSCLC) (Fennell *et al.*, 2016). Currently over 1200 clinical trials are being run for cisplatin-based combo treatments across cancers, including NSCLC (Huang et al., 2016). Thus, there is a clear effort to find adjuvant treatments for cisplatin therapy in NSCLC. Tumors develop resistance to cisplatin via a myriad of mechanisms. One key way in which they do so is through reducing DNA damage load via hijacking processes that affect the chromatin environment at cisplatin lesions (Furuta et al., 2002; Xiao et al., 2021). Thus, it is likely that chromatin states dictate cisplatin's potential to damage DNA to some degree. However, the relationship between cisplatin-induced damage and chromatin state has never been investigated in NSCLC. Here we generated a high-resolution map of DNA damage induced by cisplatin in a NSCLC model. Interestingly, we found that active and bivalent promoters are enriched for induced damage, suggesting features or processes at these loci generate either SSBs or DSBs in response to cisplatin lesions. It is possible that high levels of transcription-coupled nucleotide excision repair (TC-NER), which have been observed at promoters, could be contributing to SSB formation (Hu *et al.*, 2016; Shu *et al.*, 2016). However, transcription levels at these promoter classes do not correlate with DNA damage signal. Alternatively, it is possible that DSBs form at promoters due to proximity to efficient replication initiation sites (Langley et al., 2016; Petryk et al., 2016). Regardless, promoter-based chromatin metabolism appears to contribute to cisplatin genotoxic potential. The data generated in this study can be leveraged to further assess

chromatin-based mechanisms that contribute to cisplatin genotoxicity, with the goal of finding targetable processes to guide therapeutic development.

Acknowledgements

This work was supported by NIH grants R01 GM074701 to M.F.C., and the TRDRP 2019B Predoctoral fellowship award T30DT0906 to M.K.

Figures

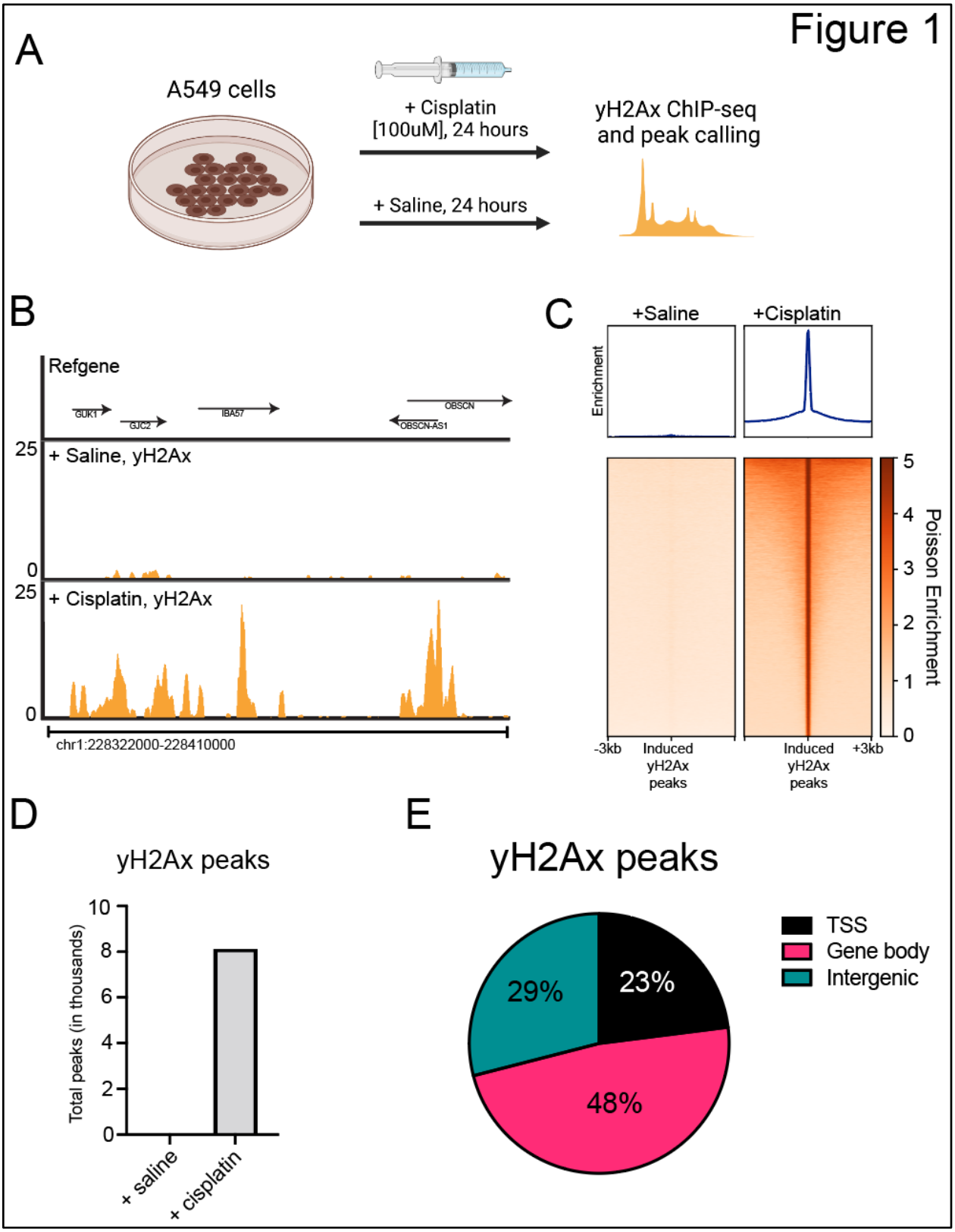


Figure 1
A
A549 cells
+ Cisplatin
[100uM], 24 hours
+ Saline, 24 hours
yH2Ax ChIP-seq
and peak calling
B
Refgene
GUK1
GJC2
IBA57
OBSCN
OBSCN-AS1
25
+ Saline, yH2Ax
0
25
+ Cisplatin, yH2Ax
0
chr1:228322000-228410000
C
+Saline
+Cisplatin
Enrichment
5
4
3
2
1
0
Poisson Enrichment
-3kb
Induced
yH2Ax
peaks
Induced
yH2Ax
peaks
+3kb
D
yH2Ax peaks
Total peaks (in thousands)
10
8
6
4
2
0
+ saline
+ cisplatin
E
yH2Ax peaks
29%
23%
48%
TSS
Gene body
Intergenic

Figure 2

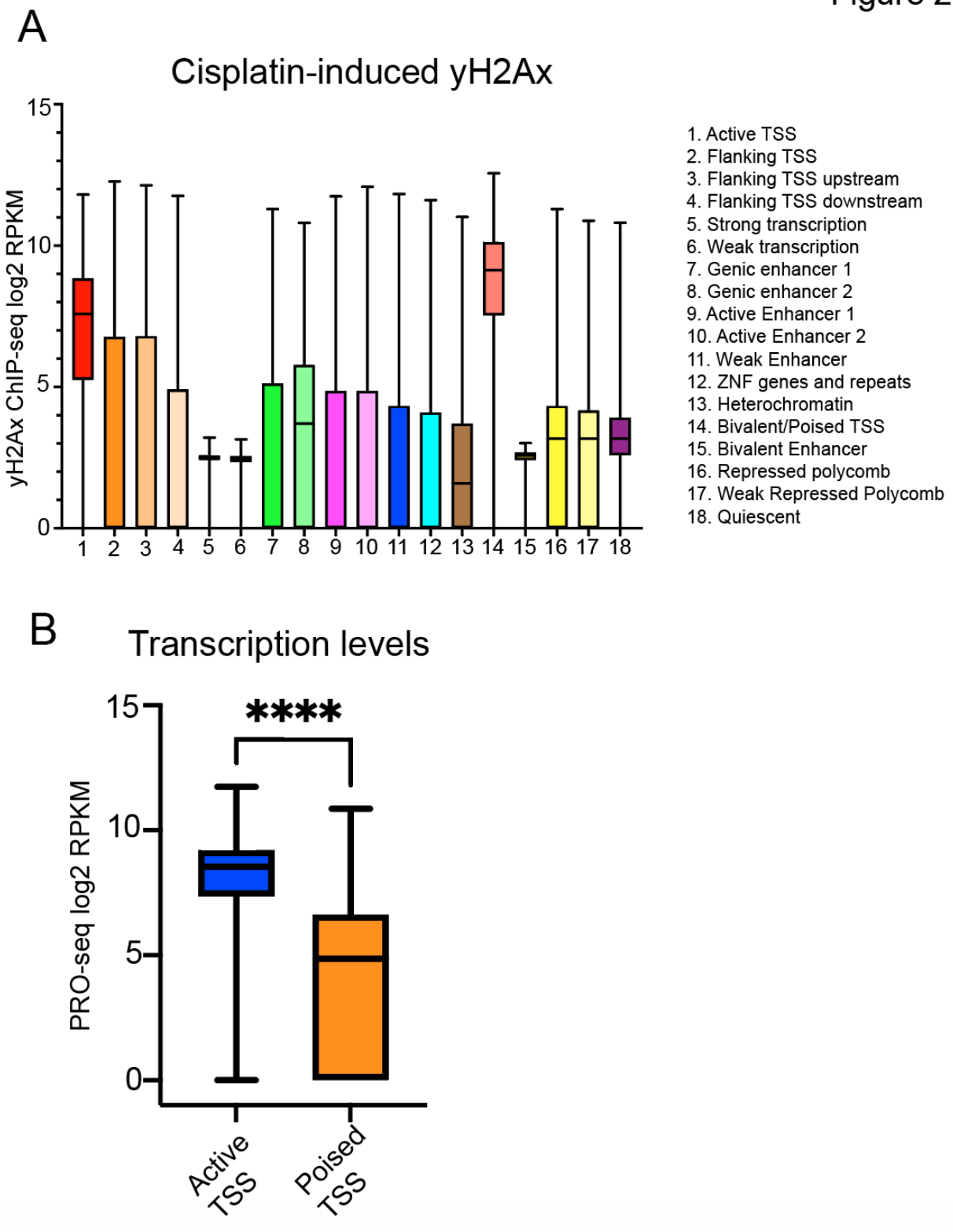

A
Cisplatin-induced yH2Ax
yH2Ax ChIP-seq log2 RPKM
15
10
5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
1. Active TSS
2. Flanking TSS
3. Flanking TSS upstream
4. Flanking TSS downstream
5. Strong transcription
6. Weak transcription
7. Genic enhancer 1
8. Genic enhancer 2
9. Active Enhancer 1
10. Active Enhancer 2
11. Weak Enhancer
12. ZNF genes and repeats
13. Heterochromatin
14. Bivalent/Poised TSS
15. Bivalent Enhancer
16. Repressed polycomb
17. Weak Repressed Polycomb
18. Quiescent
B
Transcription levels
PRO-seq log2 RPKM
15
10
5
0
****
Active TSS
Poised TSS

Figure legends

Figure 1: Genome-wide mapping of Cisplatin-induced damage across the NSCLC genome. A. Graphic representation of experimental strategy to map DNA damage induced by cisplatin in A549 cells. B. Browser track showing ɣH2Ax signal at a genomic locus in saline and cisplatin treated A549 cells. C. Average profiles and heatmaps of ɣH2Ax signal at cisplatin-induced ɣH2Ax peaks in saline and cisplatin-treated A549 cells. D. Bar chart showing the number of called ɣH2Ax peaks in saline and cisplatin treated A549 cells. E. Pie chart showing the genomic locations of ɣH2Ax peaks in cisplatin-treated A549 cells.

Figure 2: Cisplatin-induced damage across chromHMM identified chromatin states. A. Box and whiskers plot of ɣH2Ax ChIP-seq RPKMs across 18 chromatin states identified by ChromHMM. B. PRO-seq levels across active TSS and poised TSS loci.

Materials and Methods

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
yH2Ax	Abcam	Cat# ab81299
Small molecules		
Cisplatin	EMD Millipore	Cat# 232120
Kits		
KAPA HyperPrep Kit	Roche diagnostics	Cat# 07962312001
Experimental Models: Cell Lines		
A549	ATCC	CCL-185
Deposited Data		
A549 PRO-seq	John Lis, Cornell	ENCODE: doi:10.17989/ENCSR244HDV
A549 ChromHMM 18-state model	Manolis Kellis, Broad	ENCODE: doi:10.17989/ENCSR283FYU
Software and Algorithm		
Bedtools		Quinlan and Hall 2010
Samtools		Li et al., 2009

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
yH2Ax	Abcam	Cat# ab81299
Small molecules		
Cisplatin	EMD Millipore	Cat# 232120
Kits		
KAPA HyperPrep Kit	Roche diagnostics	Cat# 07962312001
MACS2		Zhang et al., 2008
Bowtie2		Langmead et al., 2009
Deeptools		Ramirez et al., 2014

EXPERIMENTAL PROCEDURES

ɣH2Ax ChIP-seq

ChIP-seq was performed in A549 cells as was done in (Xue et al., 2017)), using an antibody against ɣH2Ax. Sequencing libraries were made using the KAPA HyperPrep Kit. Fastq files were mapped to the hg19 genome with bowtie2 (Langmead et al., 2009) to generate bam alignment files. All bam files were then processed with samtools (Li et al., 2009) so that duplicates were removed, and low-quality reads were filtered out. MACS2 peakcall (Zhang et al., 2008) was then used to generate read normalized treatment and background bedgraph files from IP and input controls respectively, as well peak files. MACS2 bdgcmp (Zhang *et al.*, 2008) 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 (Consortium, 2012; Kent et al., 2010) for downstream analysis using the python deeptools software suite (Ramirez et al., 2014).

Peak subsetting by genomic feature

Peaks were intersected with either refgene TSS extended by 1kb in each direction, or refgene gene body regions with excluded TSS. Bedtools intersect was used for file processing.

ChromHMM analysis

Samtools bedcov was used to map reads from processed bam files to bed files describing loci belonging to a particular chromatin state. RPKMs per loci were then calculated and graphed via PRISM 9 software.

Cell culture

A549 cells were cultured on TC qualified plates in media containing DMEM/F12 (1:1) (Thermo Fisher, 11320-033), supplemented with 10% fetal bovine serum.

Statistical tests

P-values generated from RPKM distribution comparisons were calculated using the unpaired parametric T-test in Prism GraphPad.

References

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74. 10.1038/nature11247.

Duan, M., Ulibarri, J., Liu, K.J., and Mao, P. (2020). Role of Nucleotide Excision Repair in Cisplatin Resistance. Int J Mol Sci *21*. 10.3390/ijms21239248.

Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc *12*, 2478-2492. 10.1038/nprot.2017.124.

Fennell, D.A., Summers, Y., Cadranel, J., Benepal, T., Christoph, D.C., Lal, R., Das, M., Maxwell, F., Visseren-Grul, C., and Ferry, D. (2016). Cisplatin in the modern era: The backbone of first-line chemotherapy for non-small cell lung cancer. Cancer Treat Rev *44*, 42-50. 10.1016/j.ctrv.2016.01.003.

Frankenberg-Schwager, M., Kirchermeier, D., Greif, G., Baer, K., Becker, M., and Frankenberg, D. (2005). Cisplatin-mediated DNA double-strand breaks in replicating but not in quiescent cells of the yeast Saccharomyces cerevisiae. Toxicology *212*, 175-184. 10.1016/j.tox.2005.04.015.

Furuta, T., Ueda, T., Aune, G., Sarasin, A., Kraemer, K.H., and Pommier, Y. (2002). Transcription-coupled nucleotide excision repair as a determinant of cisplatin sensitivity of human cells. Cancer Research *62*, 4899-4902.

Hashimoto, S., Anai, H., and Hanada, K. (2016). Mechanisms of interstrand DNA crosslink repair and human disorders. Genes Environ *38*, 9. 10.1186/s41021-016-0037-9.

Hu, J., Lieb, J.D., Sancar, A., and Adar, S. (2016). Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. Proc Natl Acad Sci U S A *113*, 11507-11512. 10.1073/pnas.1614430113.

Huang, R., Langdon, S.P., Tse, M., Mullen, P., Um, I.H., Faratian, D., and Harrison, D.J. (2016). The role of HDAC2 in chromatin remodelling and response to chemotherapy in ovarian cancer. Oncotarget *7*, 4695-4711. 10.18632/oncotarget.6618.

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. 10.1093/bioinformatics/btq351.

Langley, A.R., Graf, 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. 10.1093/nar/gkw760.

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. 10.1186/gb-2009-10-3-r25.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079. 10.1093/bioinformatics/btp352.

Marti, T.M., Hefner, E., Feeney, L., Natale, V., and Cleaver, J.E. (2006). H2AX phosphorylation within the G(1) phase after UV irradiation depends on nucleotide excision repair and not DNA double-strand breaks. P Natl Acad Sci USA *103*, 9891-9896. 10.1073/pnas.0603779103.

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. 10.1038/ncomms10208.

Podhorecka, M., Skladanowski, A., and Bozko, P. (2010). H2AX Phosphorylation: Its Role in DNA Damage Response and Cancer Therapy. J Nucleic Acids *2010*. 10.4061/2010/920161.

Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res *42*, W187-191. 10.1093/nar/gku365.

Shu, X.T., Xiong, X.S., Song, J.H., He, C., and Yi, C.Q. (2016). Base-Resolution Analysis of Cisplatin-DNA Adducts at the Genome Scale. Angew Chem Int Edit *55*, 14244-14247. 10.1002/anie.201607380.

Xiao, Y., Lin, F.T., and Lin, W.C. (2021). ACTL6A promotes repair of cisplatin-induced DNA damage, a new mechanism of platinum resistance in cancer. Proc Natl Acad Sci U S A *118*. 10.1073/pnas.2015808118.

Xue, Y., Pradhan, S.K., Sun, F., Chronis, C., Tran, N., Su, T., Van, C., Vashisht, A., Wohlschlegel, J., Peterson, C.L., et al. (2017). Mot1, Ino80C, and NC2 Function Coordinately to Regulate Pervasive Transcription in Yeast and Mammals. Mol Cell *67*, 594-607 e594. 10.1016/j.molcel.2017.06.029.

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. 10.1186/gb-2008-9-9-r137.

Chapter 5: Conclusions

Transcription-replication coordination

The co-occurence of RNA transcription and DNA replication on the genome is essential for life. Both processes involve the unwinding and traversing of the DNA fiber by polymerases, potentiating transcription-replication collisions (TRCs) (Helmrich et al., 2013). The induction of TRCs in controlled systems has demonstrated that head-on collisions over R-loop forming sequences (RLFS), but not co-directional collisions, potently stall the replication fork and generate DNA breaks (Bruning and Marians, 2020; Hamperl et al., 2017; Kumar et al., 2021; Prado and Aguilera, 2005). Averting these genotoxic TRCs is thus critical for the maintenance of genome integrity and the viability of cycling cells (Hamperl *et al.*, 2017; Nojima et al., 2018). However, it has remained unclear how genotoxic TRCs are avoided on the genome. The 'passive' model proposes that replication fork movement and transcription orient co-directionally genome-wide through replication initiation site (RIS) placement upstream of active genes (Chen et al., 2019; Petryk et al., 2016). In this scenario, transcription does not need to be regulated to mitigate genotoxic TRCs, as all collisions would be co-directional, and thus tolerable, in nature. Alternatively, an 'active' model proposes that head-on transcription over RLFS does occur on the genome during the cell cycle, but is suppressed in S-phase to allow passage of the replication fork without incident. If the active model did indeed accurately describe transcription-replication coordination, then still unknown transcriptional regulatory mechanisms likely function to preserve genome integrity in human cells. In the case of tumors, these mechanisms might be highly leveraged to preserve cell viability in the face of elevated levels of replication stress (Dobbelstein and Sorensen, 2015). Thus, there is value in determining whether

transcription-replication coordination is achieved through transcriptional regulation. However, the active model has never been systematically evaluated.

In this thesis work, I leveraged a multitude of datasets to demonstrate that head-on transcription over RLFS occurs frequently on the genome, proximal to a subset of stringently selected RIS loci in both the MCF-7 and A549 cancer models. This analysis enabled the annotation of a novel class of transcriptional bodies termed head-on transcription units (HO TUs), which are pervasive and RLFS-rich in nature. Leveraging phased GRO-seq data from MCF-7 cells, I additionally find that HO TUs are downregulated during S-phase relative to the non-replicating G1-phase of the cell cycle. This observed downregulation is correlated with RLFS density. As RLFS formation post-collision drives stable fork stalling, this strongly suggests that transcription at these loci is temporally silenced to avoid genotoxic TRCs.

My analysis, although descriptive in nature, supports the active model of transcription-replication coordination. Furthermore, the unveiling of HO TUs as a novel class of transcription units provides a framework to study regulatory mechanisms that mitigate genotoxic TRCs. For example, increases in transcription generated by the inhibition of the positive elongation factors Spt6 and BRD4 in asynchronous tumor cell lines have shown phenotypes suggesting the prevention of genotoxic TRCs (Lam et al., 2020; Nojima *et al.*, 2018). However, it is unclear how these regulators affect global HO TU transcription. It is possible that different regulators silence HO TU subsets based on pervasive transcript association. The wealth of publicly available datasets can be

leveraged to investigate global HO TU regulation and identify the compendium of likely regulators. Ultimately, this work unveils a global principle of transcription-replication coordination, and provides a framework to study transcriptional regulation in the context of DNA damage prevention, especially in tumor cells.

INO80 and MOT1 in NSCLC

Non-Small Cell lung cancer (NSCLC) remains the leading cause of cancer-related death in the United States (Herbst et al., 2018). Current standards of care only extend survival on the scale of months, hinting at the need to develop more efficacious next-generation therapies (Alanazi et al., 2020; Fennell et al., 2016; Herbst *et al.*, 2018). NSCLC tumors have characteristically high levels of replication fork stalling (replication stress), as indicated by a rapid mutation rate and high ssDNA load (Boucher et al., 2019; Kandoth et al., 2013; Zhao et al., 2009). Increased ssDNA leads to depletion of the fork-stabilizing heterotrimeric complex RPA, which ablates the ability of cells to buffer further increases in stress (Toledo et al., 2017; Toledo et al., 2013). This suggests that NSCLC tumors would be sensitized to perturbations that induce an increase in stalled replication forks. Indeed, NSCLC models with loss-of-function mutations in the SWI/SNF subunit SMARCA4 have shown elevated sensitivity to CHK1 inhibition, which mechanistically catalyzes increased origin firing and the generation of stalled forks (Kurashima et al., 2020). However, CHK1 inhibition has been found to be highly toxic in the clinic (Dent, 2019). The discovery of novel mechanisms leveraged by NSCLC tumors to suppress replication stress could potentially lead to the identification of promising therapeutic targets.

INO80 is a chromatin-remodeling complex that is capable of regulating various DNA metabolic processes, including transcription (Poli et al., 2017). Past work has identified INO80 as an oncogenic factor in NSCLC (Zhang et al., 2017). However, the mechanism by which INO80 facilitates tumor growth remains unclear. Recent work in yeast and mouse embryonic stem cell models unveiled that INO80, along with the TBP antagonist MOT1, prevents genotoxic TRCs through silencing transcription at RIS (Topal et al., 2020). Interestingly, INO80 and MOT1 only prevented DNA breaks under conditions of replication stress, suggesting that an increased stress load sensitizes cells to genotoxic TRCs. Indeed, a separate study found that INO80 deletion in yeast only reduced viability in media containing hydroxyurea (Papamichos-Chronakis and Peterson, 2008). The tumor-selective toxicity observed upon INO80 inhibition in human studies suggests that the INO80/MOT1 axis might functionally prevent genotoxic TRCs, and thus replication fork stalling events, in NSCLC (Zhang *et al.*, 2017). However, this possibility has not been investigated.

In this thesis work, I show that INO80 and MOT1 bind at HO TUs in the A549 NSCLC model, and cooperatively silence transcription at a subset with elevated H2A.Z levels. Furthermore, I demonstrate that INO80 and MOT1 cooperatively suppress bulk DNA damage and facilitate tumor cell growth. DNA damage prevention by INO80 and MOT1 was found to be both replication and R-loop dependent. This collectively supports a model by which the INO80/MOT1 axis preserves the viability of NSCLC by preventing genotoxic TRCs. Interestingly, I found that INO80 inhibition generated DNA damage in A549, but not immortalized lung epithelial cells, suggesting that INO80 might be

leveraged as a DNA protectant under conditions of tumor-specific replication stress. Indeed, the A549 cell line has activating KRAS and inactivating SMARCA4 mutations, which generate replication stress phenotypes in controlled experiments.

This work elucidates an underlying mechanism that contributes to INO80's oncogenic function in NSCLC, translates a regulatory axis found in other organisms to human cells, and provides functional evidence that the silencing of HO TUs is linked to the prevention of DNA damage in tumors. These findings further support the active transcription-replication coordination model. On a large scale, the observations made in this paper support the targeting of INO80 therapeutically, possibly in combination with drugs that target replication forks. In support of this, the Rad52 inhibitor DL-DOPA has been found to selectively amplify INO80 inhibition induced apoptosis in PC3 cells (Prendergast et al., 2020). Interestingly, the metabolite Inositol-6 (IP6) directly inhibits INO80 in-vitro, and demonstrates anti-tumor effects in-vivo (El-Sherbiny et al., 2001; Shen et al., 2003; Vucenik and Shamsuddin, 2003). However, poor solubility and rapid metabolism limits the bioavailability of IP6 (Vucenik and Shamsuddin, 2003). The development of INO80-targeted small molecules with desirable pharmacokinetic and pharmacodynamic properties could potentially lead to novel therapeutics with enhanced clinical profiles relative to the current standards of care.

References

Alanazi, A., Yunusa, I., Elenizi, K., and Alzarea, A.I. (2020). Efficacy and safety of tyrosine kinase inhibitors in advanced non-small-cell lung cancer harboring epidermal growth factor receptor mutation: a network meta-analysis. Lung Cancer Manag *10*, LMT43. 10.2217/lmt-2020-0011.

Boucher, D., Ashton, N., Suraweera, A., Burgess, J., Bolderson, E., Barr, M., Gray, S., Gately, K., Adams, M., Croft, L., et al. (2019). Human single-stranded DNA protein 1 (hSSB1): a prognostic factor and target for non-small cell lung cancer (NSCLC) treatment. Lung Cancer *127*, S17-S17. Doi 10.1016/S0169-5002(19)30084-4.

Bruning, J.G., and Marians, K.J. (2020). Replisome bypass of transcription complexes and R-loops. Nucleic Acids Res *48*, 10353-10367. 10.1093/nar/gkaa741.

Chen, Y.H., Keegan, S., Kahli, M., Tonzi, P., Fenyo, 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. 10.1038/s41594-018-0171-0.

Dent, P. (2019). Investigational CHK1 inhibitors in early phase clinical trials for the treatment of cancer. Expert Opin Investig Drugs *28*, 1095-1100. 10.1080/13543784.2019.1694661.

Dobbelstein, M., and Sorensen, C.S. (2015). Exploiting replicative stress to treat cancer. Nat Rev Drug Discov *14*, 405-423. 10.1038/nrd4553.

El-Sherbiny, Y.M., Cox, M.C., Ismail, Z.A., Shamsuddin, A.M., and Vucenik, I. (2001). G0/G1 arrest and S phase inhibition of human cancer cell lines by inositol hexaphosphate (IP6). Anticancer Res *21*, 2393-2403.

Fennell, D.A., Summers, Y., Cadranel, J., Benepal, T., Christoph, D.C., Lal, R., Das, M., Maxwell, F., Visseren-Grul, C., and Ferry, D. (2016). Cisplatin in the modern era: The backbone of first-line chemotherapy for non-small cell lung cancer. Cancer Treat Rev *44*, 42-50. 10.1016/j.ctrv.2016.01.003.

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 e719. 10.1016/j.cell.2017.07.043.

Helmrich, A., Ballarino, M., Nudler, E., and Tora, L. (2013). Transcription-replication encounters, consequences and genomic instability. Nat Struct Mol Biol *20*, 412-418. 10.1038/nsmb.2543.

Herbst, R.S., Morgensztern, D., and Boshoff, C. (2018). The biology and management of non-small cell lung cancer. Nature *553*, 446-454. 10.1038/nature25183.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333-339. 10.1038/nature12634.

Kumar, C., Batra, S., Griffith, J.D., and Remus, D. (2021). The interplay of RNA:DNA hybrid structure and G-quadruplexes determines the outcome of R-loop-replisome collisions. Elife *10*. 10.7554/eLife.72286.

Kurashima, K., Kashiwagi, H., Shimomura, I., Suzuki, A., Takeshita, F., Mazevet, M., Harata, M., Yamashita, T., Yamamoto, Y., Kohno, T., and Shiotani, B. (2020). SMARCA4 deficiency-associated heterochromatin induces intrinsic DNA replication stress and susceptibility to ATR inhibition in lung adenocarcinoma. NAR Cancer *2*, zcaa005. 10.1093/narcan/zcaa005.

Lam, F.C., Kong, Y.W., Huang, Q., Vu Han, T.L., Maffa, A.D., Kasper, E.M., and Yaffe, M.B. (2020). BRD4 prevents the accumulation of R-loops and protects against transcription-replication collision events and DNA damage. Nat Commun *11*, 4083. 10.1038/s41467-020-17503-y.

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 e977. 10.1016/j.molcel.2018.10.011.

Papamichos-Chronakis, M., and Peterson, C.L. (2008). The Ino80 chromatin-remodeling enzyme regulates replisome function and stability. Nat Struct Mol Biol *15*, 338-345. 10.1038/nsmb.1413.

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. 10.1038/ncomms10208.

Poli, J., Gasser, S.M., and Papamichos-Chronakis, M. (2017). The INO80 remodeller in transcription, replication and repair. Philos Trans R Soc Lond B Biol Sci *372*. 10.1098/rstb.2016.0290.

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

Prendergast, L., McClurg, U.L., Hristova, R., Berlinguer-Palmini, R., Greener, S., Veitch, K., Hernandez, I., Pasero, P., Rico, D., Higgins, J.M.G., et al. (2020). Resolution of R-

loops by INO80 promotes DNA replication and maintains cancer cell proliferation and viability. Nat Commun *11*, 4534. 10.1038/s41467-020-18306-x.

Shen, X., Xiao, H., Ranallo, R., Wu, W.H., and Wu, C. (2003). Modulation of ATP-dependent chromatin-remodeling complexes by inositol polyphosphates. Science *299*, 112-114. 10.1126/science.1078068.

Toledo, L., Neelsen, K.J., and Lukas, J. (2017). Replication Catastrophe: When a Checkpoint Fails because of Exhaustion. Mol Cell *66*, 735-749. 10.1016/j.molcel.2017.05.001.

Toledo, L.I., Altmeyer, M., Rask, M.B., Lukas, C., Larsen, D.H., Povlsen, L.K., Bekker-Jensen, S., Mailand, N., Bartek, J., and Lukas, J. (2013). ATR prohibits replication catastrophe by preventing global exhaustion of RPA. Cell *155*, 1088-1103. 10.1016/j.cell.2013.10.043.

Topal, S., Van, C., Xue, Y., Carey, M.F., and Peterson, C.L. (2020). INO80C Remodeler Maintains Genomic Stability by Preventing Promiscuous Transcription at Replication Origins. Cell Rep *32*, 108106. 10.1016/j.celrep.2020.108106.

Vucenik, I., and Shamsuddin, A.M. (2003). Cancer inhibition by inositol hexaphosphate (IP6) and inositol: from laboratory to clinic. J Nutr *133*, 3778S-3784S. 10.1093/jn/133.11.3778S.

Zhang, S., Zhou, B., Wang, L., Li, P., Bennett, B.D., Snyder, R., Garantziotis, S., Fargo, D.C., Cox, A.D., Chen, L., and Hu, G. (2017). INO80 is required for oncogenic transcription and tumor growth in non-small cell lung cancer. Oncogene *36*, 1430-1439. 10.1038/onc.2016.311.

Zhao, Z., Xu, L., Shi, X., Tan, W., Fang, X., and Shangguan, D. (2009). Recognition of subtype non-small cell lung cancer by DNA aptamers selected from living cells. Analyst *134*, 1808-1814. 10.1039/b904476k.