Study of transcription-replication conflict and its role in genomic instability and cancer development

Yaqun Liu

▶ **To cite this version:**

Yaqun Liu. Study of transcription-replication conflict and its role in genomic instability and cancer development. Genomics [q-bio.GN]. Université Paris sciences et lettres, 2022. English. NNT : 2022UPSLS083 . tel-04298427

HAL Id: tel-04298427

https://theses.hal.science/tel-04298427

Submitted on 21 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à Institut Curie, PSL Université, Paris, France

Étude du conflit transcription-réplication et de son rôle dans l'instabilité génomique et le développement du cancer

Study of transcription-replication conflict and its role in genomic instability and cancer development

Soutenue par
Yaqun LIU
Le 27 septembre 2022

Ecole doctorale n° 515
Complexité du Vivant

Spécialité
Génetique et génomique

Composition du jury :

Allison, BARDIN Directrice de recherche, I.Curie-PSL Université	*Présidente*
Éric, LETOUZÉ Directeur de recherche, Inserm-Nantes Université	*Rapporteur*
Jean-Charles, CADORET Associate professor, IJM-Université Paris Cité	*Rapporteur*
Kathrin, MARHEINEKE Chargé de recherche, I2BC-Université Paris Saclay	*Examinatrice*
Patricia, KANNOUCHE Directrice de recherche, Gustave Roussy-Université Paris Saclay	*Examinatrice*
Philippe, PASERO Directeur de recherche, IGH-Université de Montpellier	*Examinateur*
Chun-Long, CHEN Directeur de recherche, I.Curie-PSL Université	*Directeur de thèse*

institutCurie | PSL

« From error to error, one discovers the entire truth. »

Sigmund Freud

REMERCIEMENT / ACKNOWLEDGEMENT

I would like to thank firstly all the jury members for having accepted to participate my Ph.D. defense to examinate my doctoral research work, which is my great honor to have this occasion to present what I've learnt and achieved during these previous three years in front of all of you who are absolutely brilliant experts in the fields. I will definitely get inspired from the feedbacks and exchanges with all your expertise and constructive advice and this will encourage me to be brave enough to face the further challenges. Thanks Dr. Eric Letouzé and Dr. Jean-Charles Cadoret for accepting to review my thesis and thanks Dr. Allison Bardin for accepting to be the president of my defense. Also, thanks Dr. Katherin Marheineke, Dr. Patricia Kannouch and Dr. Philippe Pasero for accepting my invitation as examinators in my defense.

Also, I would like to thank all of my thesis committee members: Dr. Axel Cournac, Dr. Gilles Fischer and Dr. Domenico Libri. Unfortunately, due to the COVID-19 sanitary situation, we were not able to organize a face-to-face physical meeting, but even in online manner, every committee meeting was very favorable with full of communication and valuable exchanges around my advancement of research work and the future career development.

Williams Yeats said, "Education is not the filling of a pail, but the lighting of a fire." I am always grateful to have met my supervisor, Dr. Chun-Long CHEN, to let me have chance to be a member of team CHEN "Replication program and genome instability" in UMR3244 "Dynamics of genetic information: fundamental bases and cancer (DIG-CANCER)", Institut Curie since my master 2 internship. Chun-Long is always able to find the best way to guild me from the very beginning till now, with proper research directions and inspiring ideas to let me free enough to explore the scientific world in the right direction. He always encourages me and provides the opportunities to participate different national/international conferences to communicate with the scientists, experts and students and trusts me to have potential in research field. Please accept my most sincere acknowledgement here Chun-Long and I am truly appreciated that he is so selfless to sparkle all of us in the lab. Then I'd like to specially thank Dr. Claude Thermes for his generous advice and encouragement during my PhD project preparation. I would like to thank Dr. Philippe Pasero and Dr. Yea-Lih Lin (IGH, Montpellier) for the super favorable collaboration with their expertise in R-loop and genomic instability field and the for the stunning results we've achieved together.

REMERCIEMENT / ACKNOWLEDGEMENT

I also would like to thank Sami who was the first team member to welcome me and help me to get familiar with the research environment in Curie. And thank all the (ex) members of the Chen's team who have been worked with me: Weitao, Inès, Stefano, Xia, Myriam, Adil, Manuela, Dalila, Joseph, Nathan, Minh-Anh, Ala Eddine and Win-Yan. I am very appreciated the memories of our regular lab meeting on Friday for exchanges/updates our work and life, of all the team building activities and unit retreats we participated, of all the help and support we gave each other. And the entire UMR3244 unit members especially Olivia, Marie-France, Rocco, Sara, Marc, thanks for all your help, support and I will cherish the happy time we spent together.

Finally, and the most importantly, I would like to thank my family especially my dear Mom for 100% supporting me, respecting all my decisions in life & study and for all the efforts my parents generously put on me. Even the current pandemic situation might make it hard to let you come to my defense physically, but I love you Mom, Dad and Sis since our hearts are always tied together.

Thanks all my relatives and friends for the supports and sweet companionship.

Cheers!

RÉSUMÉ

Au cours de chaque division cellulaire, l'ADN est un support important du matériel génétique et la réplication de l'ADN génomique est un processus essentiel pour maintenir les traits héréditaires et les activités physiologiques. Chez l'homme, chaque cycle cellulaire nécessite d'activer des dizaines de milliers d'origines de réplication pour répliquer ~ 6 milliards de paires de bases du génome afin d'assurer la transmission précise de l'information génétique. Alors que le programme de réplication est fréquemment mis à l'épreuve par le stress endogène et exogène, de nombreuses preuves au cours des dernières années ont montré que le stress de réplication induit par l'oncogène est un moteur majeur de la progression tumorale.

Les conflits entre la transcription et la réplication (TRC) surviennent parce que les machineries de réplication et de transcription partagent la même matrice d'ADN, ce qui peut se produire de manière frontale ou co-directionnelle. Les conflits frontaux sont souvent plus délétères pour conduire à l'instabilité génomique. De plus, les fourches de réplication peuvent également rencontrer des structures d'acide nucléique à trois brins appelées R-loops, qui consistent en un hybride ARN:ADN et un brin d'ADN déplacé. Cependant, le mécanisme sur la façon dont les R-loops sont impliquées dans la régulation du TRC et la stabilité génomique est encore mal connu. Ces dernières années, notre laboratoire a développé une nouvelle méthode pour mesurer directement la directionnalité de la fourche de réplication (RFD) le long du génome humain par séquençage de fragments d'Okazaki (OK-seq), qui nous fournit un outil important pour comprendre de nombreux processus liés à la réplication de l'ADN, tels que les TRC et la formation de R-loops. Pendant ce temps, malgré les protocoles bien améliorés pour OK-seq, à ce jour, il n'existe aucun outil bio-informatique disponible pour analyser les données sur RFD et détecter avec précision les zones d'initiation et de terminaison de la réplication à l'échelle du génome, ce qui rend l'étude TRC difficile à étudier.

Pour répondre à tous ces problèmes, pendant ma thèse, j'ai développé une boîte à outils bio-informatique basée sur R (OKseqHMM) pour analyser les profils RFD ainsi que pour déterminer les zones d'initiation et de terminaison de la réplication en utilisant le Modèle de Markov Caché en 4 états. Je l'ai appliqué avec succès pour analyser un grand nombre de données OK-seq et des données connexes parmi divers organismes, de la levure, de la souris aux cellules humaines. De plus, en collaboration avec le laboratoire de P. Paséro (IGH, Montpellier), nous avons réussi à montrer que les R-loops enrichies au niveau des sites de terminaison de la transcription (TTS) des gènes hautement exprimés montrent un niveau plus

RÉSUMÉ

élevé de TRC frontal. Fait important, nous avons en outre révélé que les fourches de réplication s'arrêtant à ces TTS empêchent le TRC frontal et maintiennent l'intégrité du génome d'une manière dépendante de TOP1. A part cela, nos outils et les approches d'analyse développées au cours de ma thèse peut être largement appliquée à davantage des domaines de recherche liés à la réplication, tels que l'étude de l'impact de la réplication sur la signature mutationnelle (mutations ponctuelles, variations de structure, etc.), ce qui peut contribuer une nouvelle compréhension mécanistique ainsi que le développement de nouvelles stratégies thérapeutiques contre le cancer.

MOTS CLÉS

Réplication de l'ADN, directionalité de la fourche de réplication, conflit entre réplication et transcription, R-loops, instabilité génomique, analyse des données multi-omiques

ABSTRACT

DNA is an important carrier of genetic material of cells, and the replication of genomic DNA is an essential process to maintain hereditary traits and physiological activities. In the human body, each cell division requires to activate tens of thousands of replication origins for replicating ~ 6 billion base pairs of the genome to ensure the accurate genetic information transmission. While replication program is frequently challenged by endogenous and exogenous stress, much evidence in recent years has shown that oncogene-induced replication stress is a major driver of tumor progression.

Transcription-replication conflicts (TRCs) arise because replication and transcription machineries share the same DNA template, which can occur in a head-on or co-direction manner. The head-on conflicts are often more deleterious to lead to genomic instability. In addition, replication forks may also encounter triple-stranded nucleic acid structures called R-loops, which consist of an RNA: DNA hybrid and a displaced DNA strand, that caused fork stalling even collapse. However, the mechanism about how R-loops are involved in the regulation of TRC and genomic stability is still poor known. In recent years, our laboratory has developed a new sequencing method to directly measure the replication fork directionality (RFD) along the human genome by sequencing of Okazaki fragments (OK-seq), which provides an important tool to understand many DNA replication-related processes, such as TRCs and R-loop formation. Meanwhile, despite the improved protocols for OK-seq, there was no available bioinformatics tool to analyze RFD data and accurately detect genome-wide replication initiation and termination zones, which makes the TRC study hard to investigate.

To address all these problems, during my Ph.D. study, I have developed an R-based bioinformatics toolkit (OKseqHMM) to analyze RFD profiles as well as determine replication initiation and termination zones with a 4-stage Hidden Markov Model. I have successfully applied it to analyze a large number of OK-seq and related data of various organisms from yeast, mouse to human cells. In addition, in collaboration with P. Pasero's lab (IGH, Montpellier), we successfully showed that R-loops enriched at the transcription termination sites (TTSs) of highly expressed genes showing a higher level of head-on TRC. Importantly, we further revealed that replication fork pausing at these TTSs prevents head-on TRC and maintains genome integrity in a TOP1-dependent manner. The toolkit and the analyze approaches developed during my Ph.D. study can be widely applied to other replication-related research field, such as studying the impact of replication on mutational landscape (point mutations,

ABSTRACT

structure variations, etc.), which may shed light on novel mechanistical understanding as well as the development of new therapeutic strategies for cancer.

KEYWORDS

DNA replication, replication fork direction, conflicts between replication and transcription, R-loop, genomic instability, multi-omics data analysis

ABBREVIATIONS

ACS	ARS consensus sequence
AGS	Aicardi-Goutières syndrome
ALS4	Amyotrophic lateral sclerosis type 4
AOA2	Ataxia oculomotor apraxia type 2
APOBEC	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide
ARS	Autonomously replicating sequence
AS	Ascending state
ATM	Ataxia-telangiectasia mutated
ATR	ATM and RAD3-related
bp	Base pair
BLM	Bloom syndrome RecQ like helicase
BrdU	5-bromo-2-deoxyuridine
CD	Co-direction
CDK	Cyclin-dependent kinase
CDT1	CDC10-dependent transcript 1 (CDT1)
CEAS	Cis-Regulatory Element Annotation System
CFS	Commun fragile site
CGI	CpG Island
ChIP	Chromatin immunoprecipitation
CldU	5-chloro-2'-deoxyuridine
CMG	Cdc45-MCM-GINS
CNV	Copy number variation
$CRL2^{Lrr1}$	Cullin RING Ligase 2 associated with Leucine Rich Repeats 1 [Lrr1]
CTR	Constant timing region
DDK	DBF4-dependent kinase
DDR	DNA damage response
DDX	DEAD-Box helicase

DNA	Deoxyribonucleic acid
DPE	Downstream promoter element
DRIP-seq	RNA:DNA hybrid immunoprecipitation and next-generation sequencing
dRNase H1	Defective RNase H1
DS	Descending state
DSB	Double strand break
E2	Estrogen
EdU	5-ethynyl-2’-deoxyuridine
eSPAN	Enrichment and sequencing of protein-associated nascent DNA
FACS	Fluorescence-activated cell sorter
FANCM	Fanconi anemia group M protein
FEN1	flap endonuclease 1
FRDA	Friedreich ataxia
FS	Flat state
FXS	Fragile X syndrome
G4	G-quadruplex
γ-H2AX	phosphorylation of histone variant H2AX
GLOE-seq	Genome-wide ligation of 3’-OH ends sequencing
GRO-seq	Global nuclear run-on sequencing
HO	Head-on
HMM	Hidden markov model
HU	Hydroxyurea
i-BLESS	Immobilized-Breaks Labeling, Enrichment on Streptavidin and next-generation Sequencing
IdU	5-iodo-2’-deoxyuridine
Ig	Immunoglobulin
IGV	Integrative genomics viewer
Ini-seq	Initiation sequencing
IP	Immunoprecipitation

IZ	Initiation zone
Kb	Kilobase
m6A	N6-methyladenosine
MCM	Minichromosome maintenance
mESC	Mouse embryonic stem cell
METTL3	Methyltransferase-like 3
MFI	Mean fluorescence intensity
mRNA	Messager RNA
NER	Nucleotide excision repair
NGS	Next generation sequencing
OEM	Origin efficiency metric
OH	Hydroxy
OK-seq	Okazaki fragment sequencing
ORC	Origin recognition complex
pA-Tn5	A-fused transposase Tn5
PCNA	Proliferating cell nuclear antigen
Pre-RC	Pre-replication complex
p-RPA	Phosphorylation-RPA
Pu-seq	Polymerase usage sequencing
RFD	Replication fork directionality
RNA	Ribonucleic acid
RNAP2s5	RNA polymerase 2 phosphorylated at serine 5
RNase H	Ribonuclease H
RPA	Replication protein A
RPKM	Read Per Kilobase per Million reads
RS	Replication stress
RT	Replication timing
SCAR-seq	Sister chromatids after replication sequencing
S-CDK	S phase CDK
SCF^{Dia2}	Skp, Cullin, F-box containing complex associated with Digs Into Agar 2 [Dia2]

SETX	Senataxin
SNS-seq	Small nascent strand sequencing
SSB	Single-strand break
ssDNA	Single-strand DNA
TBP	TATA-box binding protein
TdT	Terminal deoxynucleotidyl transferase
TOP	Topoisomerase
TRAEL-seq	Transferase-Activated End Ligation sequencing
TRC	Transctiption-replication conflict
TRE	Triplet repeat expansion
TRIPn-seq	Transcription-replication immunoprecipitation on nascent DNA sequencing
TSS	Transctiption start site
TTR	Transition timing regions
TTS	Transcription termination site
TZ	Termination zone

TABLE OF CONTENTS

Chapter 1 Introduction

Across species, an accurate transmission of the genomic information from parental to descendant is crucial. At each cell cycle of a human cell, tens of thousands of replication origins need to be coordinately activated to ensure the complete duplication of the about 6.4 billion base pairs (bp) in its genome. However, the DNA replication program is routinely exposed to endogenous and exogenous stresses, which play an important role in many human diseases [1]. For instance, the deregulation of this process can challenge genome stability and lead to mutations, cancers and many other genetic diseases. In particular, replication stress-induced genome alterations can represent an important early cause of cancer [2].

Replication and transcription machineries share the same DNA template then potentially interfere with each other then we called transcription-replication conflicts (TRC). They have been widely studied to be considered as an indispensable key to induce genomic instability. TRC can either be co-directional (CD) or head-on (HO). The latter has been described as more deleterious for genome integrity because of its tendency to enhance the formation of a three stranded nucleic acid structure called R-loops which consist with an DNA hybrid and a displaced DNA strand [3,4]. It is now well established that TRCs have a negative impact on genome duplication and stability, however, current evidence show that only a fraction of R-loops induce genomic instability[5]. The mechanism by which R-loops interfere with TRC enhance genome instability in mammalian cells remains poorly understood, mainly due to a lack of a comprehensive analysis of replication fork directionality.

For studying the R-loop-associated transcription-replication conflict and genome instability, we aimed to investigate (i) the whole-genome wide TRC distribution by comparing the genomic transcription direction and replication direction; (ii) the fraction of R-loop interfered with TRC that impact the replication fork progression by comparing the loci of R-loop, TRC and fork stalling; (iii) the final subset of toxic R-loops directly linked to DNA damage.

To address all these questions, the first challenge is to study the DNA replication progress especially the direction of replication fork movement. To date, more and more sequencing techniques are developed to study the replication fork progression in yeast or mammalian cells, such as :Pu-seq[6], Fork-seq[7], GLOE-seq[8], SCAR-seq[9], eSPAN[10], TrAEL-seq[11], etc. Besides, our lab has also recently developed a new method to directly measure the genome-wide replication fork directionality (RFD) by sequencing Okazaki fragments (OK-seq) [12]. In the coming sections, I will at first briefly introduce some basic knowledge on DNA replication program and then

provide detailed description for all these techniques. Finally, I will further describe the links between transcription-replication conflicts, R-loop formation and genome instability and the objectives of my Ph.D. study.

1. DNA replication

1.1. Cell cycle and checkpoints

In eukaryotes, cells followed an ordered sequence of events preparing for cell division called cell cycle. It's a four-stage process in which we observe the cell growth in size in G_1 phase, DNA duplication in S phase (DNA synthesis), preparation for division in G_2 phase and final cell segregation at mitosis (Fig.1-1). Each stage is under surveillance by the cell cycle checkpoints, which faithfully monitor and decide the timepoint to enter the corresponding phase[13]. Many proteins are involved in the control the cell cycle progression and the two principal ones are called cyclins and cyclin-dependent kinase (CDK). The kinase activity of a CDK depends on the interaction with a cyclin partner and that cyclins are tightly regulated. In G_1 phase, cells begin to grow and produce the nutrients and essential elements for cell proliferation. Normally somatic cells are supposed to have a longer G1 phase (6-12h) to ensure replication origin licensing and to guarantee complete DNA duplication [14]. It should be noted that, if G_1 gets shortening in somatic cells, for instance, by overexpressing cyclin E, to alter the normal G_1-S transition, it will deregulate the replication fork progression and induce DNA damage[2]. Thus, G_1 checkpoint is the most decisive point for a cell, at which it must choose whether to divide or not depends on how the cell cycle goes. Either cell may leave the cell cycle and enter a resting state called G_0 phase waiting for resuming the cell cycle process, or the cell passes the G_1 checkpoint, enters S phase and it becomes irreversibly committed to cell division. Here Cyclin D-CDK4 and Cyclin E-CDK2 complexes are formed to phosphorate protein Rb to promote the cell enter S phase for the further DNA replication[15].

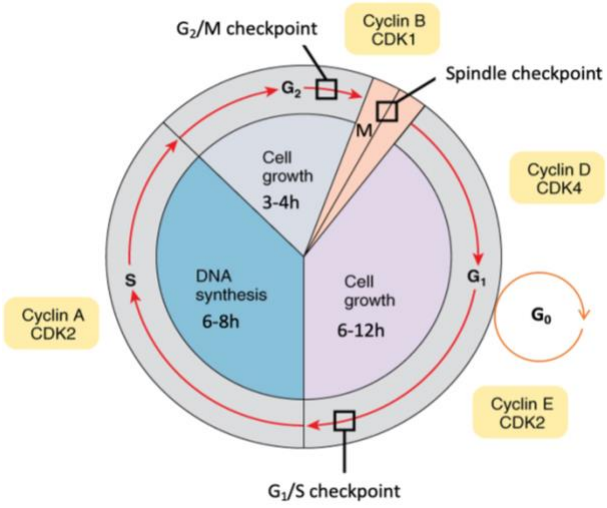

Figure 1-1. A typical cell cycle of eukaryotic cells.

Once passed the G1 check point, cells enter S phase, Cyclin A is produced and complex with CDK2 to activate DNA replication program. To make sure that cell division goes smoothly, there is a G_2/M checkpoint before mitosis to check the DNA integrity and DNA replication completeness [16,17] . If it detects incomplete DNA synthesis or replication errors/damage, the cell will pause at the G_2/M checkpoint to allow to either complete DNA replication or repair the damaged DNA. If the damage is irreparable, the cell may undergo apoptosis to ensure that damaged DNA is not passed on to daughter cells and is important in preventing cancer. The last step, cell enters mitosis and comes to M checkpoint, also known as the spindle checkpoint, to verify if all the sister chromatids are correctly attached to the spindle microtubules to prepare anaphase chromosome segregation, which is activated by Cyclin B-CDK1 complex [16,17].

Each phase of the cell cycle is well controlled and highly surveyed by checkpoints that ensure the correct replication and segregation of the genetic information into daughter cells. This is extremely important because any replication error will be transmitted to subsequent generations where it could contribute to the development of cancer in muti-cellular eukaryotes.

1.2. Replication process

1.2.1. Initiation

In each cell cycle of eukaryote cells, replication initiation events are widely activated along the genome. The loci from which replication starts are called replication origins. Although the basic replication machineries are very conserve in all eukaryotes, replication initiation mechanisms are species-specific for the evolutionary adaptation. For instance, in budding yeast *Saccharomyces cerevisiae*, specific short replicator sequences called autonomously replicating sequences (ARSs) represent potential origins and it's the only eukaryote in which the origin recognition complex (ORC) can recognize a clear AT-rich consensus sequence around 17 bp named ACS (ARS consensus sequence) for DNA replication [18]. However, in metazoans, we were not yet able to identify clear ARS since much more origins are activated with more variable DNA replicating features.

During the G_1 phase, origins are marked by the formation of a pre-replicative complex (pre-RC). Its assembly starts with the binding of ORC1-6 to DNA and it is followed by the recruitment of cell division control protein 6 (CDC6), chromatin licensing and DNA replication factor 1 (CDT1) and the minichromosome maintenance (MCM) helicase double hexamer complex, which contains the six subunits MCM2–7 [19,20]. Pre-RC assembly then "licenses" the

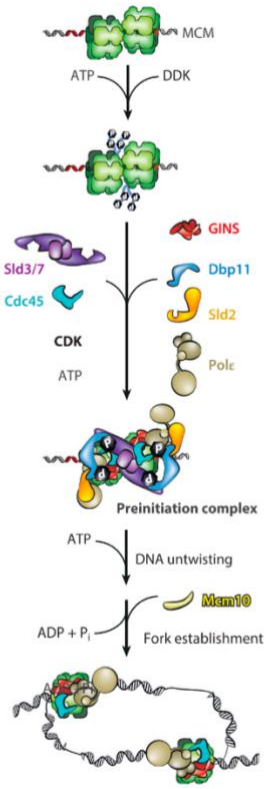

Figure 1-2. Initiation of eukaryotic DNA replication. DNA-loaded DHs are phosphorylated by DDK. Phospho-Mcm4 and -Mcm6 are recognized by Sld3, which exists in complex with homo-dimeric Sld7 and recruits Cdc45 onto MCM. CDK targets Sld3 as well as Sld2. Phospho-Sld2 and phospho-Sld3 bind to Dpb11, which also binds Pol ε and GINS. Sld2, Dpb11, GINS, Pol ε, Sld3/7, and Cdc45 binding to a DH forms the preinitiation complex. Release of ADP and binding of ATP leads to stable CMG formation, concomitant with DH interface disruption and origin DNA untwisting. Addition of Mcm10 switches on ATPase-powered DNA unwinding by MCM, causing two CMG particles to cross their paths, which establishes bidirectional replication forks. CDK, cyclin dependent kinase; CMG, Cdc45–MCM–GINS; DDK, Dbf4 dependent kinase; DH, double hexamer; Mcm, minichromosome maintenance; Pol ε,DNApolymerase ε. Figure adapted from Costa (2022).

origin for potential activation in the subsequent S phase. During initiation, the DNA structure is made differentially accessible to the proteins and enzymes involved in the replication process. As CDT1 is recruited by CDC6 while it is bound to MCM2-7 and CDK1 can regulate replication by inhibiting MCM loading via phosphorylating ORC and activating helicase. When the levels of CDK1arise, CDT1 is degraded therefore MCM2-7 cannot be recruited anymore at the origins. Consequently, MCM loading can only occur during G1 phase when CDK activity is low, and origins can only fire after G1 phase when CDK levels rise. CDKs and DBF4-dependent kinases (DDKs) also help to recruit corresponding proteins, including CDC45 and GINS, at the origin sites to form the CDC45/MCM2–7/GINS (CMG) helicase complex to initiate DNA unwinding, facilitate formation of the replisome, and prime DNA synthesis. MCM10 is another key helicase activator, which can strongly bind to the CMG complex in S phase for unwinding the DNA and recruiting the polymerase α [21]. SLD2 and SLD3 which are the two key CDK substrates, and DPB11 are also recruited for helicase activation [22–25]. Furthermore, three DNA polymerases mainly participate in eukaryotic replication: Polymerase α, ε, and δ (Pol α, ε, and δ) [26]. Proliferating Cell Nuclear Antigen (PCNA) is recruited to form a homo-trimeric ring-shaped sliding clamp for tethering Pol ε and δ (Fig. 1-2).

Each helicase unwinds and separates the double DNA helix into two single-stranded DNA. As the DNA opens up, Y-shaped structures called replication fork are formed. Replication Protein A (RPA) binds to both strands to promote replication fork stabilization and DNA repair [27]. Since two helicases bind, two replication forks are formed and are extended in both directions

as replication proceeds creating a replication bubble and we say a replisome is assembled at the replication origin. There are multiple origins of replication on the eukaryotic chromosome, which allow replication to occur simultaneously in hundreds to thousands of locations along each chromosome. Not all of them are activated in each cell division. When the cells are under replicative stress, these backup origins can be licensed by the excess MCM2-7 guarantee the whole replication program to be completed [28].

1.2.2. Elongation

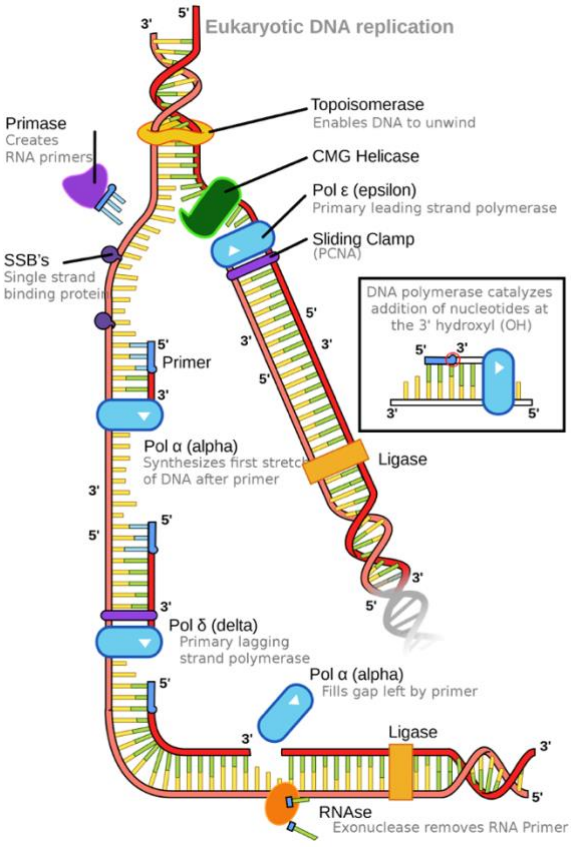

Figure 1-3. Eukaryotic DNA replication process on both DNA strands. Helicase opens DNA double strands. The primase creates short RNA primers on DNA to initiate the synthesis of the and polymerases. The replication is only in the 5'-3 direction, the leading strand is replicated continuously and the lagging strand in a discontinuous way with formation of Okazaki fragments. Figure adapted from Wikipedia (https://commons.wikimedia.org/wiki/File:Eukaryotic_DNA_replication.svg).

DNA polymerase cannot initiate directly new strand synthesis and it only adds new DNA nucleotides to the 3′ end of the newly synthesized polynucleotide strand. All new strands must be initiated by a specialized RNA polymerase called primase. Primase initiates polynucleotide synthesis and by creating a short RNA polynucleotide strand complementary to template DNA strand, which is called the primer. Once RNA primer has been synthesized at the template DNA, primase exits, and DNA polymerase extends the new strand with nucleotides (A, T, C, or G) complementary to the template DNA [26]. DNA polymerase can only synthesize new strands in the 5′ to 3′ direction. Therefore, the two newly synthesized strands proceed in opposite directions because the template strands at each replication fork are complementary, which lead to as we called the semi-discontinuous mechanism of DNA replication. The "leading strand" is synthesized continuously in the same direction as the growing replication fork whereas the "lagging strand" is synthesized in the direction away from the replication fork. This lagging strand is synthesized fragmentarily and therefore has to constantly encounter the previously new synthesized DNA sequences. These

new short pieces, each around 150 nucleotides in length in eukaryotes, are called Okazaki fragments and each fragment begins with its own RNA primer [29] (Fig. 1-3). During the fork progression, unwinding of the parental double DNA helix by DNA helicases locally generate compensatory positive torsional stress that can result either supercoiling ahead of the fork or pre-catenanes between two replicated duplexes behind the fork by interlocking the DNA molecules at the fork branch. Positive supercoils can be removed in eukaryotes by Topoisomerase type IB (TOP1) and type IIA (TOP2), furthermore TOP2 is required for the chromosomal decatenation, which maintain all along the progression of the replication [27].

1.2.3. Termination

Since pairs of replication forks that assemble at thousands of replication origins almost simultaneously and then move in opposite directions, DNA replication finishes when converging replication forks meet. DNA synthesis is completed and we therefore called replication termination. Once all the template nucleotides have been replicated, the replication process is not yet over. RNA primers need to be replaced with DNA nucleotides by proteins FEN1 (flap endonuclease 1) and RNase H. The enzymes FEN1 and RNase H remove RNA primers at the start of each leading strand and at the start of each Okazaki fragment, leaving gaps of unreplicated template DNA then gaps are connected rapidly by ligases. Later in S phase, replisomes encounter each other when it

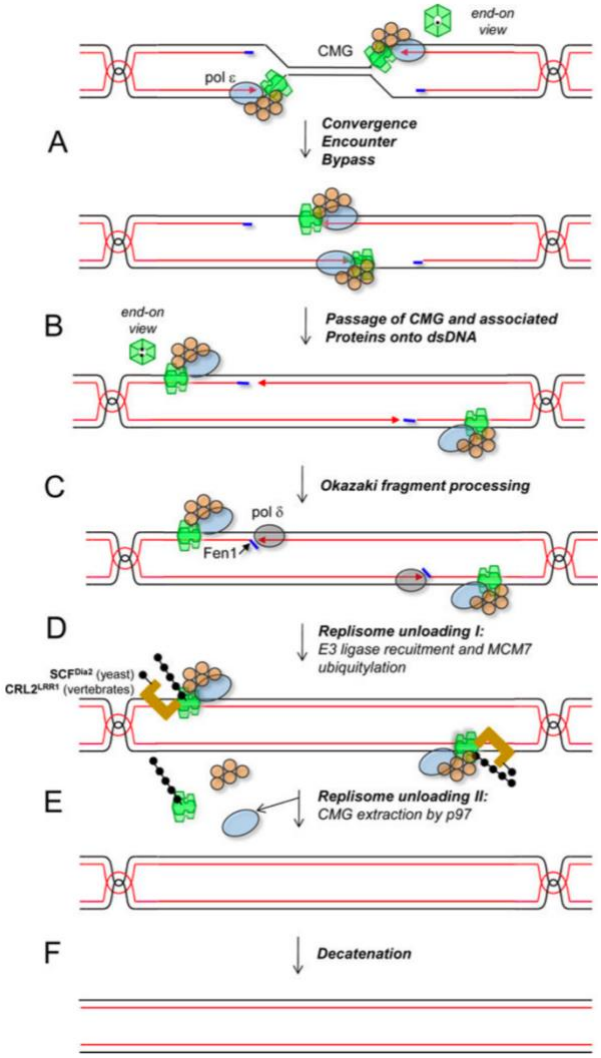

Figure 1-4. Eukaryotic DNA replication termination. (A) In late S phase, forks come too close to each other to allow formation of supercoils in the unreplicated DNA, leading to the onset of convergence. During convergence, which lasts until forks encounter each other, topological stress is relieved by the formation of pre-catenanes. An end-on view of CMG illustrates the presence of single-stranded DNA in its central channel. (B) The encounter causes no detectable fork stalling, implying that converging CMGs bypass each other. After bypass, CMG helicases keep translocating until they reach a downstream Okazaki fragment. (C) The CMG helicases pass over the ssDNA–dsDNA junction and keep moving on dsDNA (see end-on view). (D) The leading strand is extended to the downstream Okazaki fragment. The last Okazaki fragment is processed, possibly by de novo recruitment of DNA Pol δ and by 3' flap processing by flap endonuclease 1 (FEN1). (E) Once CMG encircles dsDNA, it undergoes polyubiquitylation on its MCM7 subunit by SCF^{Dia2} or $CRL2^{Lrr1}$. The ubiquitylated MCM7 is extracted from chromatin by the ATPase p97. (F) Catenanes are removed. Figure adapted from Dewar et al. 2017

reaches to DNA template that has already been replicated to lead fork convergence. The converging CMG complexes encounter each other on different strands since CMG interact mainly with leading strand therefore facilitating bypass without pausing [30]. The CMG helicase dissociation linked with MCM7 ubiquitylation is the following key event in termination process which require the E3 ubiquitin ligase SCF^{Dia2} in yeast (Skp, Cullin, F-box containing complex associated with Digs Into Agar 2 [Dia2]) or $CRL2^{Lrr1}$ (Cullin RING Ligase 2 associated with Leucine Rich Repeats 1 [Lrr1]) in vertebrates [30]. The other proteins without interaction with CGM are eventually removed independently of replication termination. Chromosomal decatenation is carried out by TOP2 and the replication process is finally completed (Fig. 1-4).

1.3. Replication timing and relative detection techniques

Not all the replication origins are fired at the same time during S phase but they follow a strict temporal program with each chromosome containing segments that are replicated towards the beginning of the S phase (early replicated domains) or the end of it (late replicating domains), which is referred to as the replication timing (RT) program [31]. RT is conserved among eukaryotes but also cell-type specific, and correlated with many epigenomic features. Early replicating domains are located in the nuclear interior and they are enriched in active histone modifications. Late replicating domains are localized at nuclear and nucleolar periphery and are enriched in repressive histone markers [32]. Recent research has indicated that the RT program represents a very stable epigenetic feature of chromosome: most expressed genes generally reside in the early-replicating regions while late-replicating loci are mostly transcription silencing, less structured and have elevated mutation rates in the human germ line, in somatic cells and also in cancer cells [31,33]. Aberrant replication program is associated with changes in gene expression, changes in epigenetic modifications and an increased frequency of structural rearrangements [33].

Up to now, the RT profile, describing how to derive information from raw experimental data of mammalian cells, now is well developed and can be easily measured. One of the primary methodologies is to use nucleotide analogs 5-bromo-2-deoxyuridine (BrdU) to pulse-label newly synthesized DNA during S phase. Then BrdU-labeled cells are sorted directly into S or G1 phase just based on the copy numbers of DNA by fluorescence-activated cell sorter (FACS). Another similar method used the same labeling step but FACS sorted into early and late S-phase populations based on DNA content of each cell then the labeled DNA is immune-precipitated with an anti-BrdU antibody (BrdU-IP), and DNA synthesized either early or late

is determined by microarray or next generation sequencing (NGS) [34] (Fig.1-5A). After mapping to genome, we can calculate the read counts of log2(S/G1) or log2(S_{early}/S_{late}) in a defined window size to get the primary RT profiles [35]. However, based the technical limitation, a high ratio of noise background effects the results since the log2 strategy cannot really identify the true nascent DNA sequences. Ever since, the higher resolutive and well improved technique is Repli-seq to determine the accurate genome-wide replication timing in mammalian cells. Starting from E/L Repli-seq [36] which only fractionates the early/late replicating domains to multi-fraction Repli-seq using 4-6 fractions of S phase [37–39]. The primary RT profiles showed us large constant timing regions (CTRs) and timing transition regions (TTRs), the resolution of these profiles does not allow us to identify initiations and termination events. RT profiles are also cell type specific with characteristic replication patterns that reveal a significant replication plasticity covering > 50% of the human genome [37] (Fig. 1-5B,C).

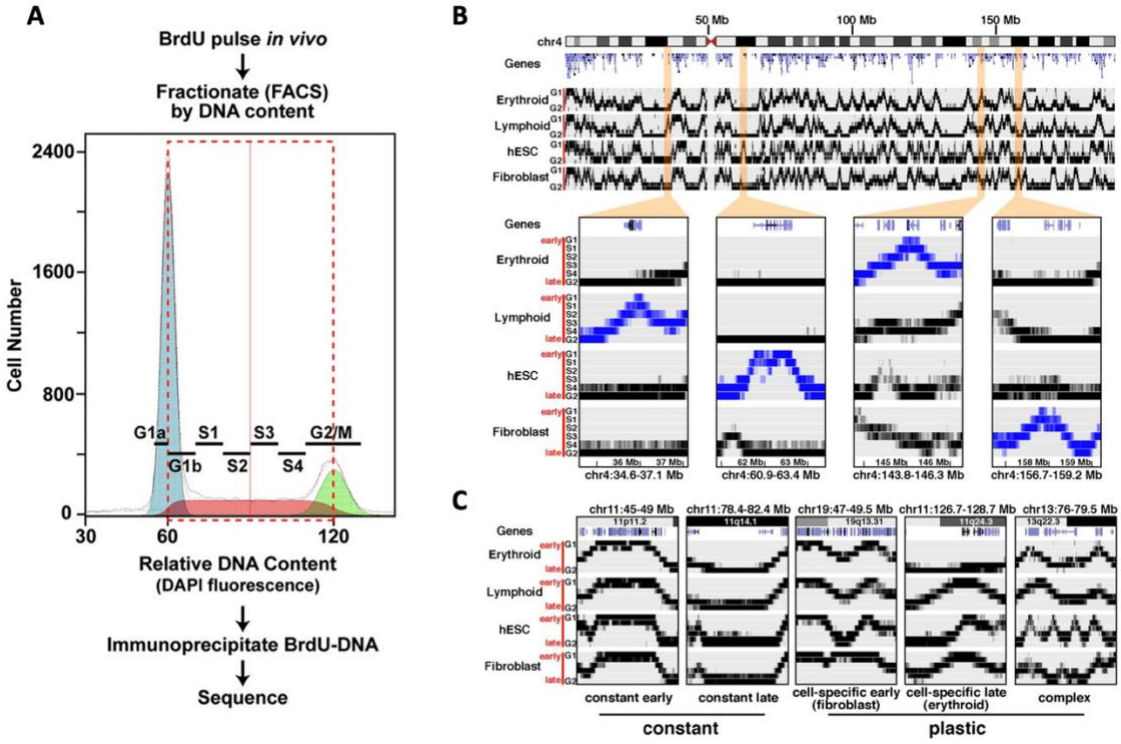

Figure 1-5. Replication timing profiling. (A) Cell-cycle fractionation of newly synthesized DNA. Exponentially growing cells are pulse-labeled with BrdU, stained with DAPI, and sorted into different fractions of the cell cycle according to DNA content as shown for this normal lymphoblastoid cell line (LCL). Fractionation is continuous across the cell cycle. Antibody-purified BrdU-labeled DNA is made into sequencing libraries and sequenced on the Illumina platform, and the sequence tags are mapped to the human hg18 reference genome. (B) Comparison of replication timing profiles from four cell types across chromosome 4, illustrating unique lineage patterns. Lineage-specific early-replication patterns. Cell-lineage-specific early patterns are highlighted in expanded chromosome 4 subregions. The lymphoid-specific CENTD1 gene in the 34.6–37.1 Mb region is at the apex of an early-replication peak in the GM06990 LCL, whereas this region is uniformly late-replicating in the other three lineages. Similar patterns are seen in the other expanded regions: LPHN3 in the 60.9–63.4 Mb region (hESC-specific), GYPA-GYPB-GYPE in the 143.8–146.3 Mb region (erythroid-specific), and PDGFC in the 156.7–159.2 region (fibroblast-specific). (C) Stereotypical replication timing patterns. Shown are major patterns of DNA replication timing observed across the genome, including (i) regions of constant early replication across cell lineages; (ii) regions of constant late replication; (iii) regions with cell-specific early replication; (iv) regions with lineage-specific late replication (i.e., one cell type late, all others early); and (v) complex patterns that vary considerably between lineages. Adapted from Hansen et al. 2009

The most recent Repli-seq extends to single-cell level based on the copy number variation (CNV) to try to decipher the replication timing variation in cell-to-cell[40]. Then it further improved the accuracy of RT profiling within 16 S phase fractions and is sensitive enough to detect replication initiation zones (IZs), termination regions (TZs), late CTRs and TTRs in 50kb resolution in mammalian cells [41]. However, to investigate the heterogeneous replication initiation mechanism with at least 50% tissue-specific variation RT, many approaches, which are more specific to the replication initiation and replication fork progression, are developed in recent decades.

1.4. Techniques for replication initiation detection

- **DNA combing**

DNA combing is the first single-molecule method applied to replication origin detection since 1997 [42]. It is based on the labeling of newly synthesized DNA by two sequential pulse of 5-iodo-2'-deoxyuridine (IdU) and 5-Chloro-2'-deoxyuridine (CIdU), two thymidine analogs, in an asynchronous cell population. DNA fibers are then stained with YOYO-1 and stretched on silanized glasses in pH 5.7 can be visualized by immunofluorescence with a conventional optical microscope (Fig. 1-6) [43,44]. This is a global method to determine the real spacing of origins along DNA. However, it is not suitable to identify a given origin DNA sequence due to the lack of corresponding genomic reference and the low probability to specifically target the region while the replication fork is passing through the origin.

In practice, this method can work for sequences that are hundred-fold repeated in the genome and is able to reflect the variation of velocity of replication fork progression under different conditions. In addition, one limitation that DNA combing is restricted to the DNA length which it was not able to get molecules longer than 600-800 kb [18], while it has been overcome and successfully extend the combing method into megabase level [45]. Although mapping of replication origins is more challenging, DNA combing and related fiber assay become now a golden standard to study replication fork speed in cells under normal growth or under various replication stress.

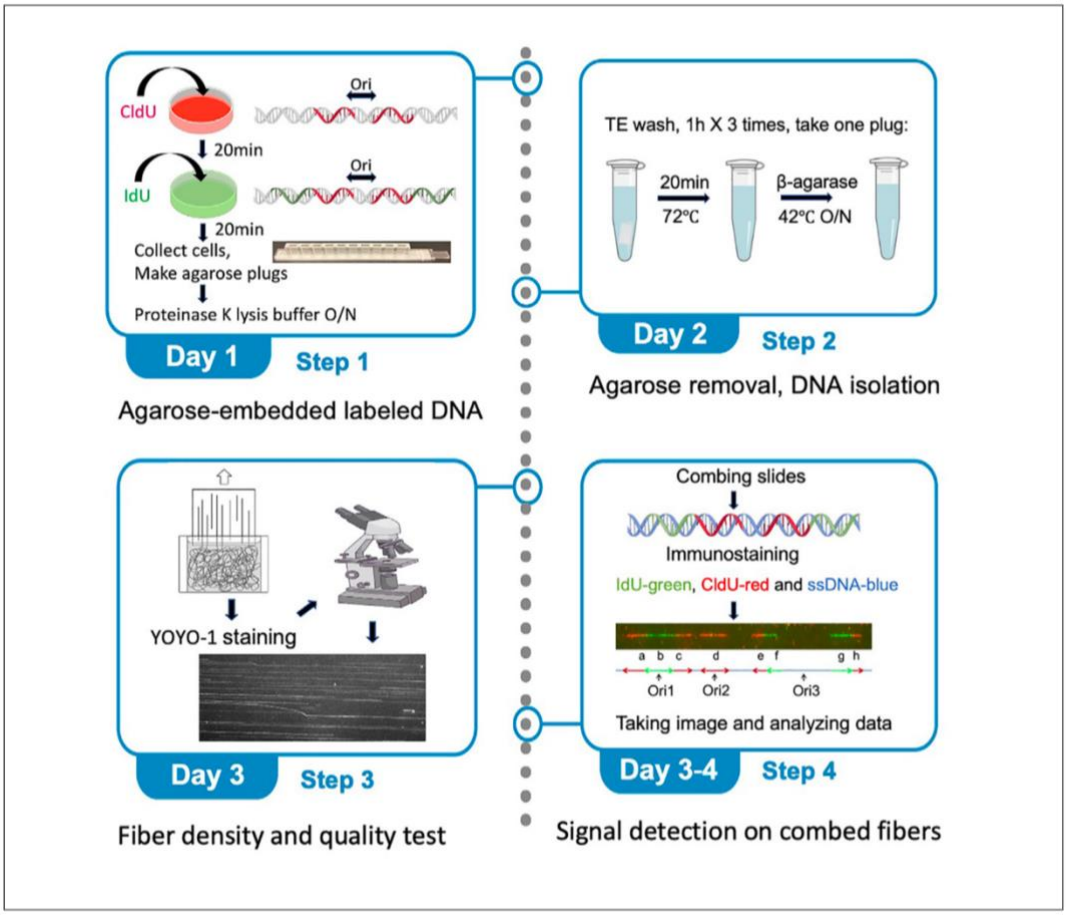

Figure 1-6. protocol of DNA replication profiling by molecular combing. Cells are proceeding the IdU and CIdU labeling then get stained with YOYO1 before DNA combing. Figure adapted from Fu et al. 2022.

- **Nanopore sequencing**

The Nanopore sequencing is a single molecule technology. A DNA filament is pushed through a membrane channel and the instrument measures the current passing through it. This current changes based on the base composition of the segment crossing the nanopore channel allowing to reconstruct the DNA sequence of a certain filament [46].

Based on Nanopore sequencing, D-NAscent technique (Detecting Nucleotide Analogue signal currents on extremely long nanopore traces) was developed. This approach is based on the possibility to distinguish between thymidine and BrdU in a DNA molecule. BrdU can therefore be used to label newly synthetised DNA in synchronized cells to extract information about the active replication origins sites, fork direction, termination sites, and fork pausing/stalling events [46]. Based on this approach, the same authors have developed other two techniques: Fork-seq, which applies the nanopore sequencing in yeast to map the replication in 200 nucleotide

resolution directly from asynchronous growing cells [7] and NanoForkSpeed, which succeeded to map and extract the velocity of individual forks in asynchronously growing cells [47].

- **SNS-seq**

As described in the previous chapter, when an origins is activated, it produced short RNA:DNA hybrids from which replication can proceed bidirectionally. Such RNA:DNA stretches are called short nascent strand (SNS) and they are used by the SNS-seq to map origins of replication.

To isolate SNS we need to remove from our samples genomic DNA that can be intact or fragmented and Okazaki fragment. To remove gDNA and Okazaki fragment we run the samples on sucrose gradient gels and extract from the gel fragments between 0.5-2.5 kb (gDNA is much longer than 2.5kb and Okazaki fragment are usually 150-200 bp long). The recovered fraction still contains SNS and fragments of broken gDNA. To get rid of the latter, samples are treated with an excess of λ-exonuclease, a 5'->3' DNA specific exonuclease that will digest the broken gDNA but not the SNS that are protected in position 5' by an RNA primer [48–50]. Another alternative protocol is to incorporate BrdU to label the nascent DNA and size-fractionate to get the BrdU-labeled SNS following by BrdU-immunoprecipitation (BrdU-IP) [51].

Although SNS-seq should provide a high-resolutive replication initiation mapping, there is always a debate about the delicate step of purification of SNS, which needs to precisely control the size-fractionation to exclude the background DNA sequences and Okazaki fragments. Besides, the two methods all have concerns which the λ-exonuclease needs to be well controlled and BrdU-labeled method may generate short DNA sequences by breakage mixed with the real SNS. Moreover, this technique, because of its nature, is limited to the identification of very efficient, well localized origins of replication.

- **Bubble-trap (Bubble-seq)**

Once the replication initiates, DNA helicase unwinds the DNA double helix and form transient bubble-shaped structure centered around an origin of replication. Bubble-seq (Bubble-trap) is devised for isolating these circular DNA fragments that contain replication initiation sites (bubbles) by following steps: DNA is digest using the restriction enzyme EcoRI. Depending on relative position of a replication fork and a EcoRI recognition sequence fragment can have different shapes: linear, Y-shaped (replication forks) and O-shaped (replication bubble

containing the two divergent forks) and the other non-DNA materials [52]. Fragments are then separated by electrophoresis on an agarose gel. Since the migration in native condition is influenced by the shape of the fragment, linear sequences are the fastest to pass through the agarose gel, followed by Y-shaped ones while the replication bubbles have relatively the slowest speed and after a long electrophoresis are the only ones still present in the gel. Captured bubbles can then be recovered and used to make libraries (Fig. 1-7) [35,53]. Bubble-seq can detect both efficient and inefficient origins with larger initiation size, which make less overlapped (< 45%) with the initiation sites that SNS-seq detected [50]. Besides, the purification of the real replication bubbles still needs to be improved since some larger-size Y-shaped fragments can also mix into the final filtration with bubbles.

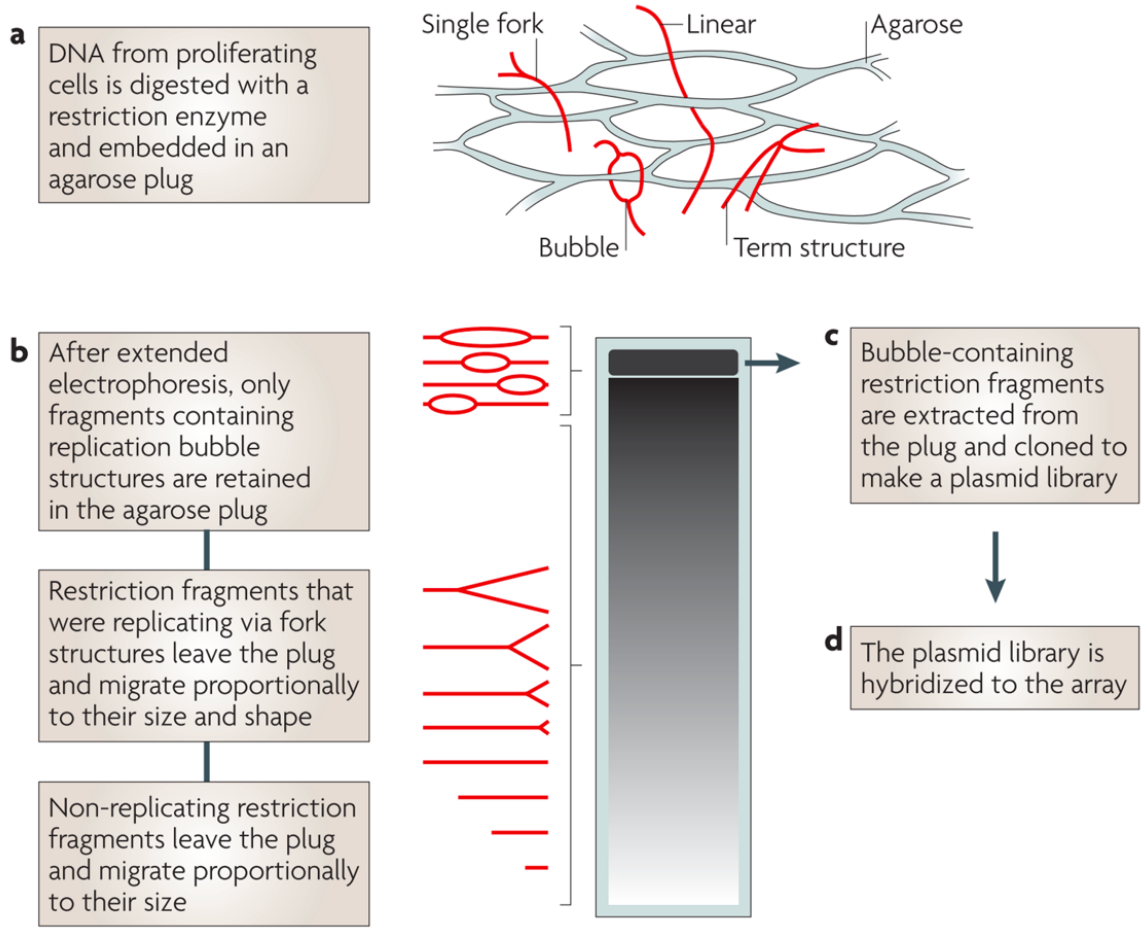

Figure 1-7. Schema of Bubble-trap protocol. Details shown in the figure. Figure adapted from D. Gilbert 2010

- **OK-seq**

Since Okazaki fragments (discontinuous short nascent fragments) are only generated on lagging strand during DNA synthesis, it provides a proper way to get the replication direction information and the initiation loci. The enrichment of Okazaki fragments for direct sequencing was first achieved in *S. cerevisiae* through ligase and checkpoint inactivation in 2012 [54]. Then OK-seq is well established to quantify the replication fork directionality (RFD) and accurate initiation and termination zones genome wide in mammalian cells [12]. Nascent DNA sequences are labelled by EdU then fractionated into pieces. Since the average size of RNA-primed Okazaki fragments are around 150-200 bp, Only the sequences that are less than 200 nucleotides are selected. To isolate the EdU-labeled replicated DNA from the RNA sequences and the short DNA sequences generated from the size-fraction step, EdU is coupled with biotin in click reaction. The biotinylated fragments are captured by streptavidin before PCR amplification and sequencing (Fig. 1-8).

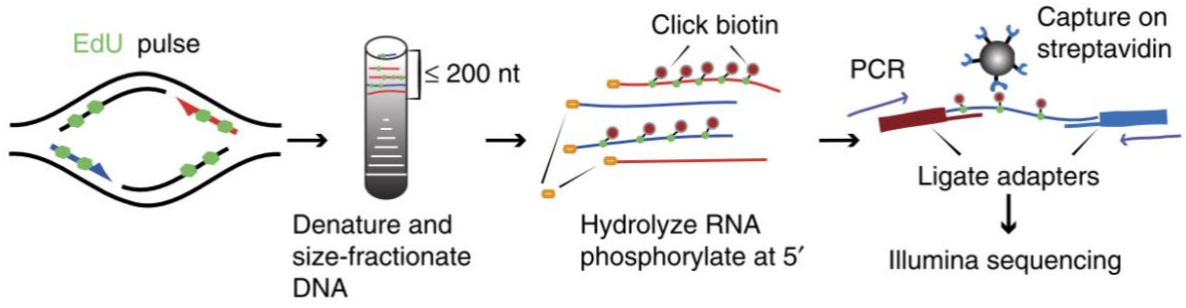

Figure 1-8. OK-seq protocol. Cells are pulsed with EdU labeling then size- fractionated to pick fragments with less than 200 bp. All the DNA fragments are proceeded with biotinylation and captured the nascent DNAs on Streptavidin beads following by sequencing. Figure adapted from Petryk et al 2016.

Okazaki fragments mapping to the Watson (+) and Crick (-) strands are generated by leftward- (L) and rightward- (R) moving forks, respectively can be normalized to $RFD = (C - W)/(C + W)$ to get RFD profile (Fig. 1-9A), and the genomic zone associated with a positive slope is considered as replication initiation zone (IZ) since the fork direction of two sides is divergent (Fig. 1-9B), while the genomic zone associated with a negative slope is referred to as termination zone (TZ) since the fork direction of two sides is convergent (Fig. 1-9C). The corresponding amplitude indicates the fire efficiency of each IZ/TZ in cell population level. With the bioinformatic method I developed, more than 10, 000 IZs can be identified with an average length of 20-30 kb for human cells. The concrete bioinformatics analysis is described in *Chapter 2 OKseqHMM*.

OK-seq is therefore a unique method to detect the replication initiation events by targeting the center of two divergent Okazaki fragments movement. However, OK-seq also exists its own limitations. As any cell population method, OK-seq averages the cell-to-cell variability. Like most of replication-related technology, OK-seq requires an enormous amount of starting material to have enough cells in S phase since the half-life of Okazaki fragments is very short e.g., GM06990 cells required 8-10 × 10^8 cells per biological replicate [12]. The initiation events detection method based on the RFD transition profiles could also be an issue since not all of the IZs are efficient enough to generate a positive slope of RFD, such like the late replicating regions with much more random initiation origins firing, which make the flatter RFD profile and are finally ignored by the detection.

Please check details with the fork direction information in the next *session 1.5*.

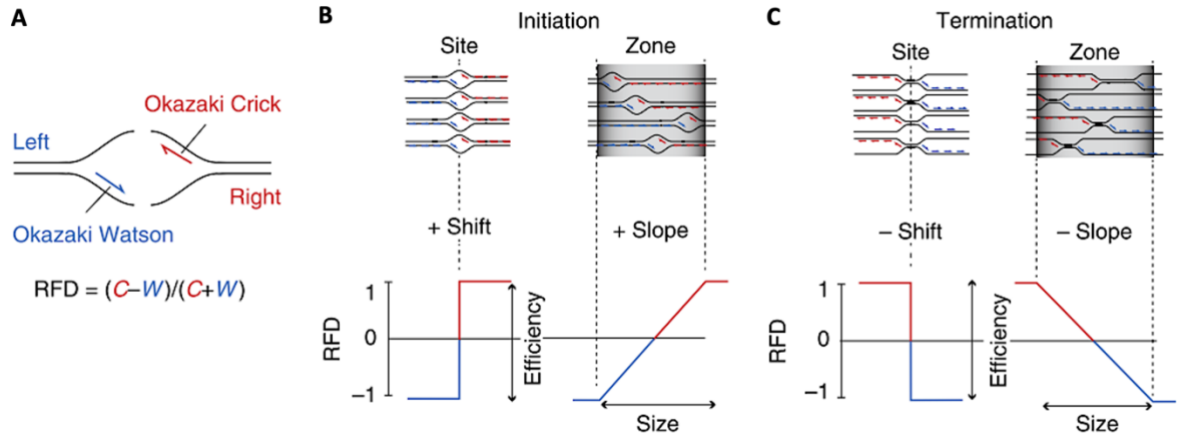

Figure 1-9. Schema of computing RFD mapped from OK-seq. C: Crick strand; W: Watson strand. Figure adapted from Petryk et al. 2016.

- **ORC/MCM ChIP-seq**

During G1 phase, cells start recruiting the ORC, CDC6, CDT1 and MCM2-7 to form the pre-replication complex to license the origins or initiate the DNA replication in S phase. Therefore, the distribution of ORC/MCM can also effectively reflect the potential replication initiation events in a whole genome scale. Since there are excessive MCMs recruited before the replication initiation, targeting MCM could detect both efficient active replication origins and the dormant or less efficient origins. Researchers have been working on this to identify these replication initiators from yeast to high eukaryotic cells and confirmed the colocalization of MCM and ORC especially in budding yeast. Hence, basically using the anti-ORC and anti-MCM antibodies then ChIP-chip (Chromatin immunoprecipitation-microarray) or ChIP-seq

(Chromatin immunoprecipitation-sequencing) [55] to obtain the ORC-/MCM- binding sites [35,56]. These sites are then compared with other technologies such as SNS-seq, OK-seq and Repli-seq to classify the active initiation sites and dormant origins. A recent evidence also confirmed that ORCs/MCMs are abundantly enriched at gene promoters/enhancers and more in early replicating domains but they don't have an obvious role into the regulation of replication initiation firing probability in human cells [57]. Nevertheless, since the replication initiation event is well agreed with a stochastic firing mechanism, the initiation could occur a bit distantly from the ORC binding sites. And ORC/MCM sites have less sequence specificity in human cells, which means it might bind to different sites in different cells [35]. In addition, the detection of both active origins and dormant ones is a strength of technique without doubt, but it's also a limitation since there is no functional method to directly distinguish the two clusters to date.

- **EdU-seq-HU**

BrdU and EdU are the widest used thymidine analogs to label nascent DNA sequences in most of the initiation detection technologies. Low concentration of EdU is sufficient and easier to obtain the incorporation signals with nascent DNAs since it can be detected in double-stranded DNA and the corresponding fluorescent label is small permeable molecule to access the DNA with usage of antibody compared with BrdU which can be only detected in single-stranded DNA condition [58]. EdU is however toxic to cells to potentially arrest the cell cycle and have to proceed the labeling procedure more than one cell cycle. Hydroxyurea (HU) is normally treated to cells to slow down the fork progression with a low concentration and it eventually facilitates the EdU-incorporation. EdU-seq-HU method is therefore come out to detect the early replication initiation events. Cells are firstly synchronized in entry of S phase by flow cytometry then are released in EdU and HU contained medium, followed by the biotin click and streptavidin capture to obtain the EdU-bind nascent DNA sequences [59]. This method restricts on the very early stage of S phase with synchronized cell population for which the detection of origins is quite limited. Besides, cell synchronization and the treatment of EdU and HU also make the procedure more complexed and more potential to get cells contaminated.

- **Ini-seq**

Alternatively, ini-seq is another independent sequencing technique to map the human DNA replication origins in genome-wide. Cells are firstly synchronized in late G1 phase of which

the replication is initiated by a cell-free DNA replication initiation assay and incorporating digoxigenin-11-dUTP for 15 minutes incubation. The newly replicated DNA labeled with digoxigenin-11-dUTP are sheared by sonication into 100-1000 bp fragments following by immunoprecipitation, PCR and sequencing (Fig. 1-10) [60]. More than 25,000 discrete replication origins are identified by this method with a reasonable consistent with other technologies like SNS-seq and OK-seq. The cell-free system used in this method truly facilitated the nascent DNA labeling step but the incubation of this specific nucleotide analog dUTP might potentially impact the results or decrease the resolution of mapping.

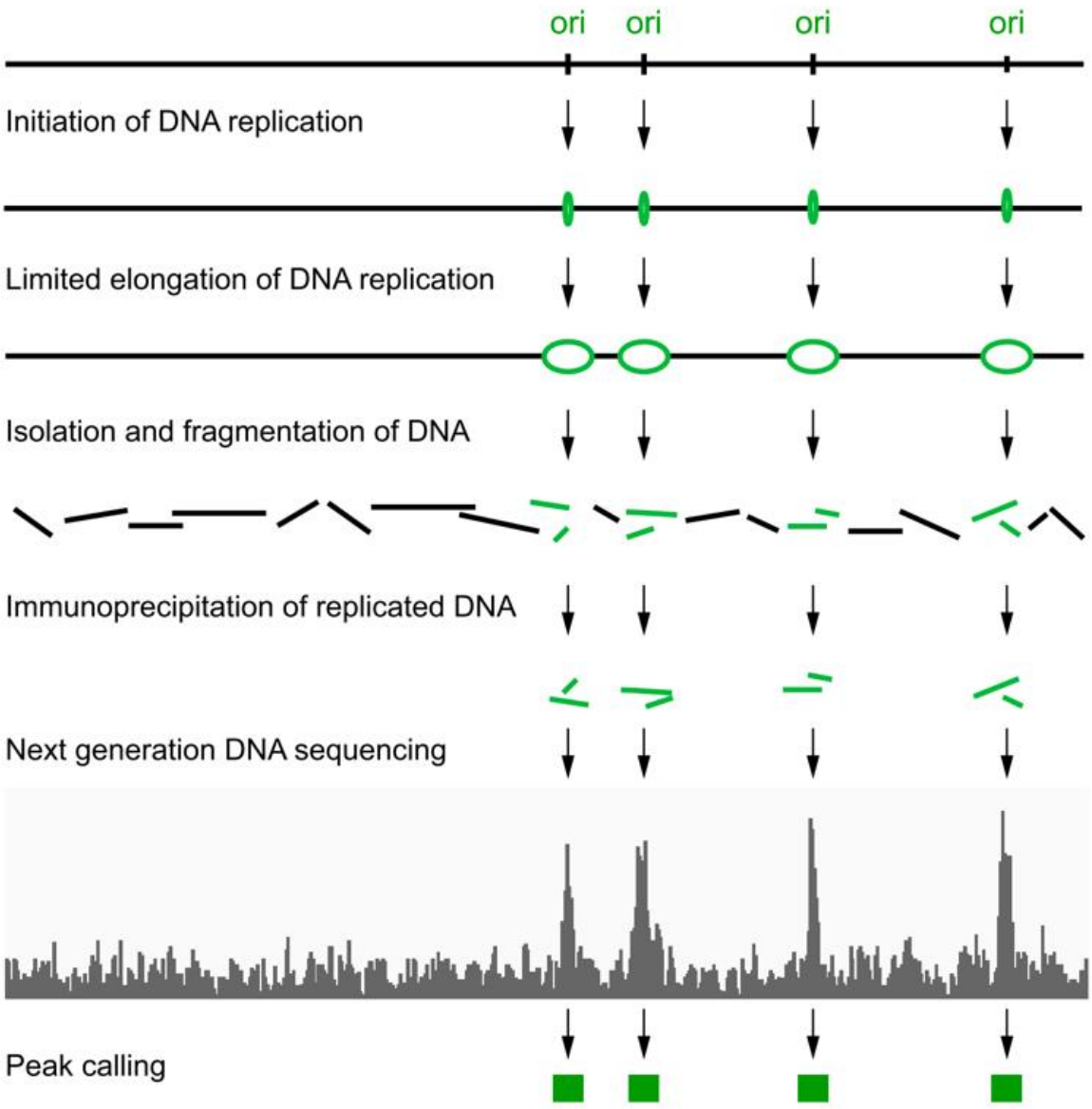

Figure 1-10. Schema of ini-seq to get the replication initiation sites. Figure adapted from Langley et al. 2016

- **ORM (2021)**

As mentioned above, almost every technique has its own advantages and disadvantages. To date, more and more novel techniques in single-molecule or single-cell level have been developed to optimize the detection of initiation sites and relative replication firing time. Optical Replication Mapping (ORM) single-molecule technique using Bionano high-throughput imaging approach is well established to investigate both spatial and temporal distribution of replication origins by mapping the long individual DNA molecules [61]. Cells are synchronized at G1/S transition using Aphidicolin, then labeling newly synthetized DNA is labeled with Aminoallyl-dUTP-ATTO-647N (a red fluorophore) at different timepoints after release in S-phase (5, 10, 20, 30, 45, 60, or 90 minutes) to observe the movement of the replication forks through ORM signals at the different time points after S phase entrance.

To identify the genomic position of a fork, ORM relies on the mapping approach of the bionano system that is based on the labeling with a green fluorophore of recurrent sequences in order to create unique spatial patterns such as NLRS (Nick, Label, Repair, and Stains) or DLS (Direct Label and Stain) and the DNA fibers are maked in blue. DNA samples are then loaded on the Saphyr chip where they are stretched inside nanochannel arrays through electrophoresis. DNA fibers can finally be accurately measured by Bionano imaging method (Fig. 1-11A) [61].

This technique generates more than 27 million DNA fibers with average length of 300 kb allowing identifying the early initiation sites and corresponding firing probability with high accuracy. By comparing with different origin-detecting methods, ORM replication tracks correlate better with Ini-seq ($r = 0.59$), followed with OK-seq ($r = 0.49$), less well with SNS-seq and Orc1 ChIP-seq (Fig. 1-11B). Their results also confirmed that the replication initiation is a stochastic mechanism conserved from yeast to humans. Besides that, ORM can map the ongoing replication fork direction in asynchronous cells, which is well corelated with the RFD profiles of OK-seq [61]. However, ORM also has their own technical limitations on the choice of thymidine analogs for the nucleoside-labeling step since neither BrdU nor EdU are compatible with Bionano technique.

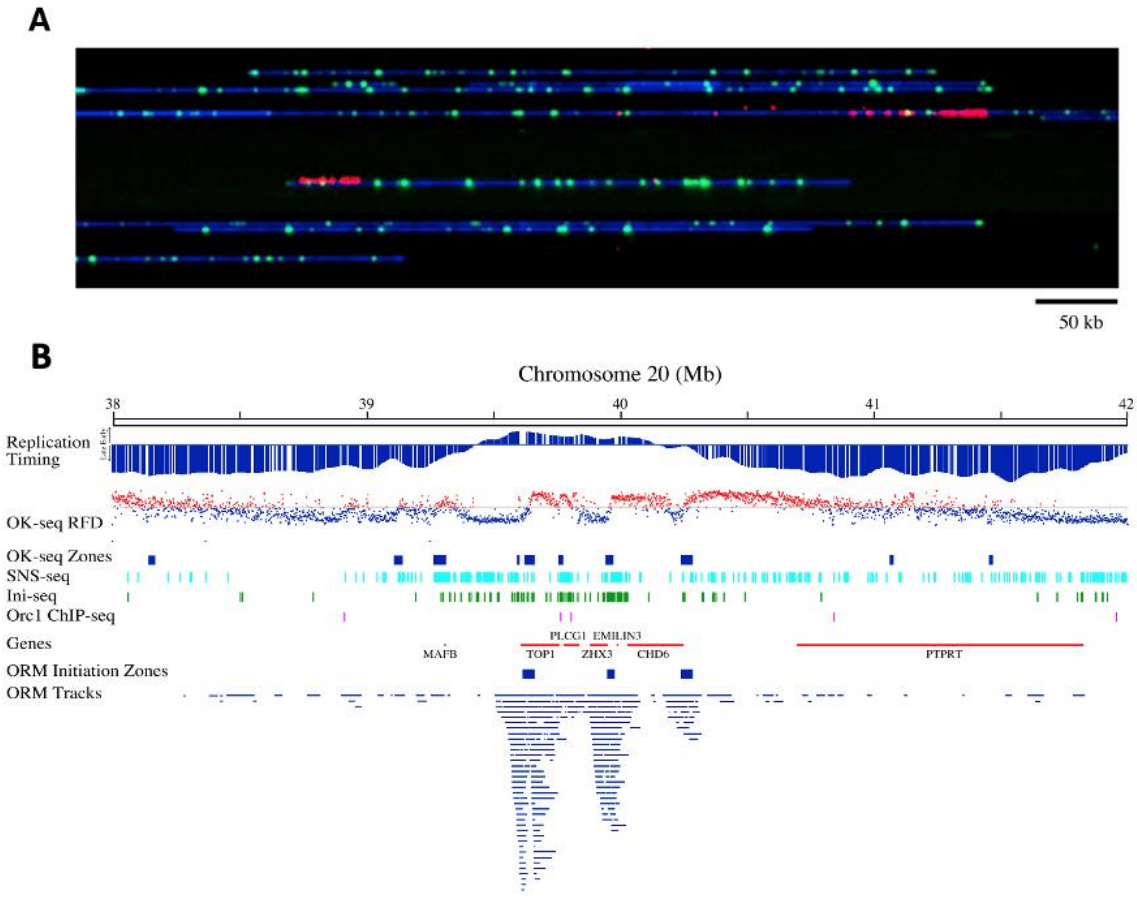

Figure 1-11. DNA fibers detected by ORM and comparison among different techniques. (A). Image of the Bionano data with DNA in blue, Nt.BspQI (NLRS motif) restriction sites in green, and incorporated fluo-dUTP-labeled early firing sequences in red. The field of view is 750 kb. (B) Comparison of ORM tracks, ORM IZs, OK-seq RFD, OK-seq IZs, SNS-seq, Ini-seq and ORC1 ChIP-seq at the TOP1 site**.** Figure adapted from Wang et al. 2021.

1.5. Detection of replication fork direction techniques

- **N-domains/nucleotide composition skew**

In prokaryote genomes, replication related DNA strand asymmetry patterns (i.e., G ≠ C and T ≠ A then we called TA and GC skews) or nucleotide compositional asymmetries, which are gradually formed along the genomic mutation events in long-term evolution, are discovered and established. It brings an original concept to detect not only the origins, but also arise a challenge to be able to capture the replication fork orientation genome-wide [62–65].

TA skew of a defined window (i.e., 1kb) is calculated as $S_{TA} = (T - A)/(T + A)$ and GC skew as $S_{GC} = (G - C)/(G + C)$, and the total skew as $S = S_{TA} + S_{GC}$ [59]. The genome-wide analysis of these skew profiles along the human genome brought to the identification of N-shape domains (termed N-domains) that correspond to a cascade of replication origins initiated at

highly efficient initiation zones locate at the borders (Fig. 1-11A, the vertical black lines) and the replication transition between two neighbor origins resulting in N-shape reflecting the mean replication fork direction, which are perfectly matched with the U-shape domains derived from replication timing profiles (Fig. 1-11B) [64,65]. The overlap between the borders of N-domains and the peaks of replication timing profiles also validated the hypothesis that these borders zones termed S-jumps (sharp positive slopes mathematically) are associated with replication initiation zones [64]. The discovery of nucleotide compositional skew concept fundamentally inspired the whole replication research field for developing techniques to detect the replication origins and fork orientation information.

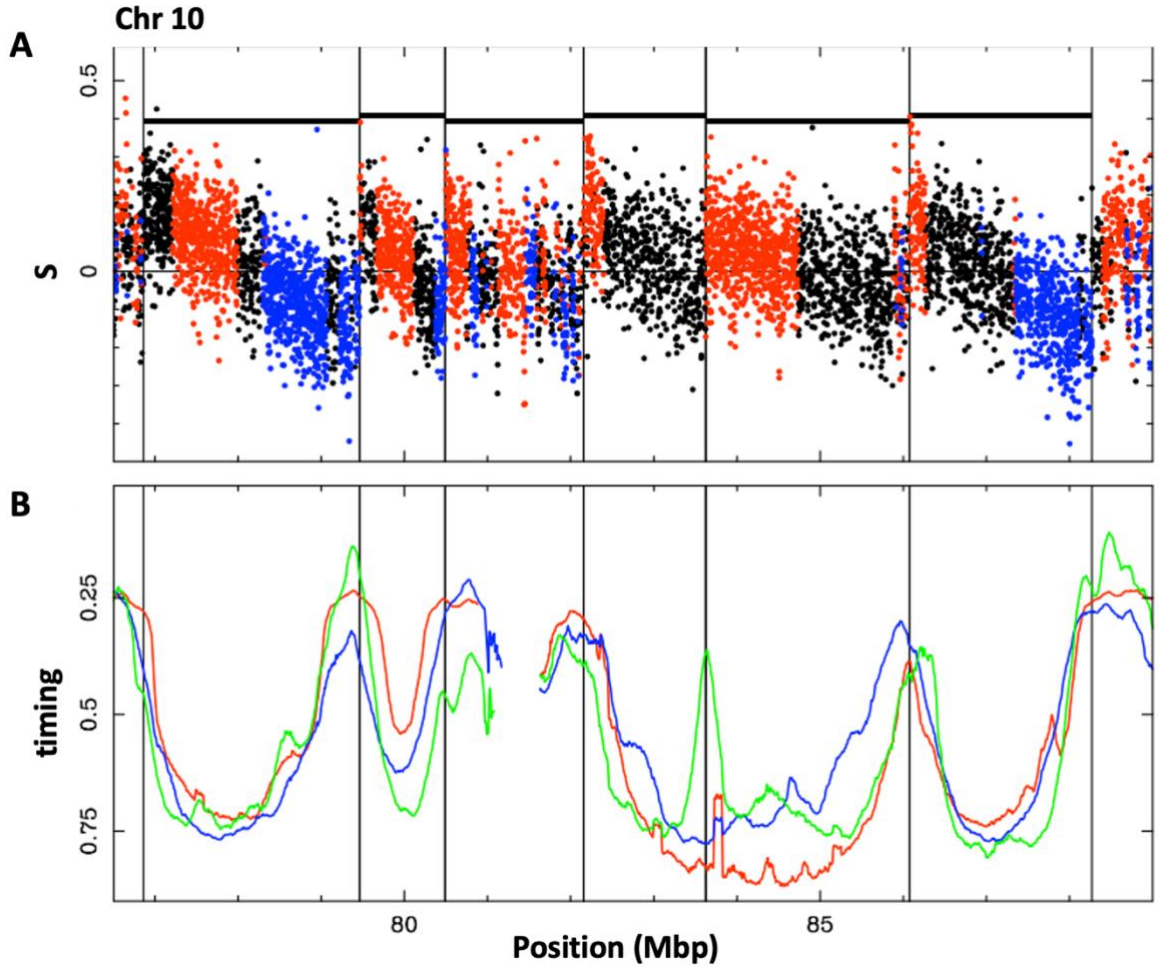

Figure 1-11. Compositional skew and replication timing profiles. (A) Skew profiles $S = S_{TA} + S_{GC}$ calculated in 1-kb windows of the repeat-masked sequence. Horizontal black lines mark the replication N-domains. Vertical lines mark the corresponding putative replication initiation zones. Black, intergenic regions; red, sense (+) genes; blue, antisense (−) genes. (B) Mean replication timing determined in BG02 ESC (green), K562 (red) and GM06990 (blue) cell lines. Figure adapted from Hyrien et al. 2013.

- **PU-seq**

DNA Polymerases are recruited differently on the two replication strands: Pol α-primase initiates replication then is rapidly replaced by elongation Pol ε on the leading strand and Pol δ on lagging strand. Polymerase usage sequencing (PU-seq, also known as HydEn-seq [66]) has been developed firstly in yeast to determine the distribution of embedded nucleotides that can be able to detect, not only the replication origins sites, but also the origin firing efficiency, estimation of replication timing and termination information, as well as the fork progression direction [6]. Ribonucleotides are alkali-labile so that it can cause strand fragmentation under alkali treatment. To map the polymerase usage, two polymerase mutations are generated respectively in RnaseH2-deficient (RnaseH2 is the main ribonuclease to remove ribonucleotide from duplex DNA by ribonucleotide excision repair) cells with alkali treatment: cdc20-M630F (Pol ε) and cdc6-L591G (Pol δ) following by sequencing. The relative ratio of read counts from Pol ε and Pol δ datasets was calculated to provide the frequency of both of polymerases used on Watson and Crick strand and direct measure of the proportion of replication fork movement leftward and right ward across the genome [6]. Therefore, the transition of polymerase usage is able to determine the initiation sites, termination sites and also the RFD. Meanwhile the sharpest changes of ratios reflect the most efficient origins. Recent updated Pu-seq technique is successfully extended to human cells [67].

- **OK-seq**

Following by the same concept as N-domain theory above-mentioned, OK-seq is established to provide the efficiency of origin usage, replication fork progression and termination information which fully extend the dimension of the initiation detection approache. Protocol and brief data analysis already mentioned in the previous session *1.4*.

By calculating the read counts of Okazaki fragments on Watson and Crick strand respectively into RFD profile (normalised range from -1 to 1), it directly demonstrates the replication fork direction, initiation zones (IZs, positive slopes) and termination information (negative slopes) with corresponding efficiency of origins usage (sharpness or amplitude for each slope) across the whole genome (Fig. 1-12).

OK-seq is the first full-scale and well-developed technique to delineate the high-resolutive human replication fork direction landscape in genome-wide with at least 10,000 accurate IZs/TZs are detected with corresponding origin usage efficiency by directly sequencing

Okazaki fragments to ensure to get the true fork movement information compared with the other techniques. With the complete bio-informatic pipeline that I developed, to date we can optimise the data analysis up to 1 kb high resolution in human and even up to 500 bp resolution in yeast. Details of the tool performance are described in *Chapter 2*. The abundant replication-related information contained by OK-seq provides an indispensable source to study the DNA replication and genomic instability research.

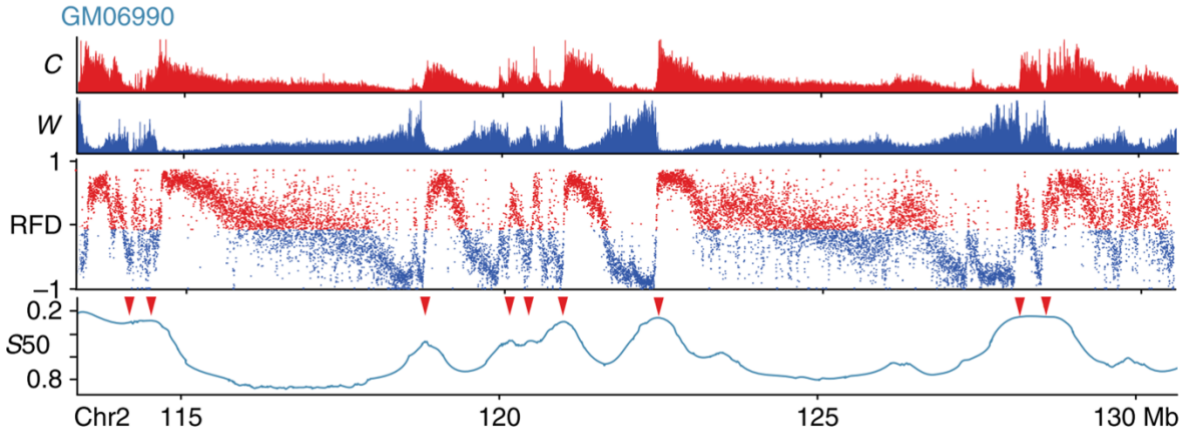

Figure 1-12. Replication fork direction obtained by OK-seq. C: Crick strand (red). W: Watson strand (blue). RFD profile calculated in 1kb adjacent binsize. S50: replication timing profile. Red arrows: early replication peaks with S50 < 0.5. Profiles are captured in chromosome 2 of GM06990 cell line. Figure adapted from Petryk et al. 2016.

- **GLOE-seq**

The GLOE-seq method is based on genome-wide ligation of 3'-hydroxy (OH) ends followed by sequencing, which maps single-strand breaks (SSB) in a strand-specific manner, to reveal the distribution of spontaneous SSBs that are distinct from double-strand breaks (DSB) patterns. Besides, it can also map Okazaki fragments escaping the size selection step in OK-seq to provide fine-resolution RFD profiles in yeast and human cells [8].

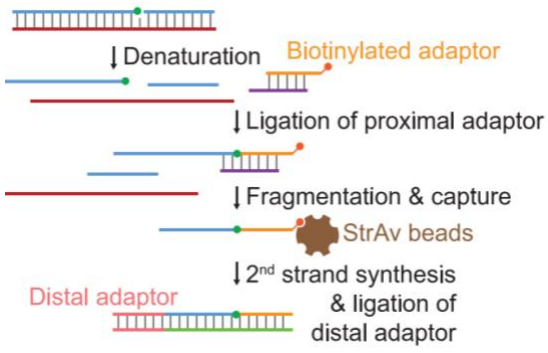

Figure 1-13. GLOE-seq workflow. green dots: ligatable 3'-OH terminus; red dots: biotin. Figure adapted from Sriramachandran et al 2020.

Genomic DNA sequences isolated from around million cells are embedded in agarose. The first step consists in the heat-denaturation of the samples that brings to the formation of ssDNA with/without SSB containing an available 3'-OH. A biotinylated adaptor is then ligated to the exposed 3'-OH. Samples are then sonicated to obtain fragments of 200nt. Biotinylated ssDNA fragments can then be captured using streptavidin

beads. Using a primer complementary to the biotinylated adaptor the 2nd strand of each ssDNA is synthesized with blunt end and a distal adaptor ligation reaction. Libraries are then amplified and sequenced (Fig. 1-13) [8].

This technology can detect Okazaki fragments as well and it has been used on human cells. To this purpose the experiment is performed in a depletion of ligase 1 and 3 depleted background, therefore impairing the ligation of adjacent fragments [8].

GLOE-seq, as any method for mapping SSBs, has high background and the the signal-to-noise ratio needs to be imporved in the future. Compared to OK-seq, GLOE-seq truly uses a reduced input cell number (~ 700,000 human cells) and avoids the size selection and analog labelling step, however, it requires replicative ligases inactivation which might be a potential issue to ensure whether the ligation is delicately inhibited and could lead to potential DNA damage response that disturb the investigation.

- **TrAEL-seq**

Like GLOE-seq, Transferase-Activated End Ligation sequencing (TrAEL-seq) is an alternative technique to map DNA 3' ends at double-strand breaks (DSBs) undergoing the DNA resection and indirectly provides RFD information in yeast and human cells.

Basically, agarose-embedded genomic DNAs are incorporated with a terminal deoxynucleotidyl transferase (TdT) that can add 1 to 4 adenosine nucleotides onto single-stranded DNA 3'ends, forming a substrate for DNA adaptor ligation by RNA ligases. Two TrAEL-seq adaptors are constructed: adaptor 1 is a hairpin to convert single-stranded ligation products into double-stranded DNA incorporated with biotin for the further purification. Once the adaptor 1 is ligated, a thermophilic polymerase with strong strand displacement and reverse transcriptase activities (e.g., Bst2.0 polymerase) extends the hairpin to form unnicked double-stranded DNA, following by sonication into fragments and the biotinylated adaptor-ligated fragments are purified on streptavidin beads. During fragmentation, DNA ends are generated and are ligated to adaptor 2. Both of adaptors are cleaved before the library amplification and sequencing (Fig. 1-14) [11].

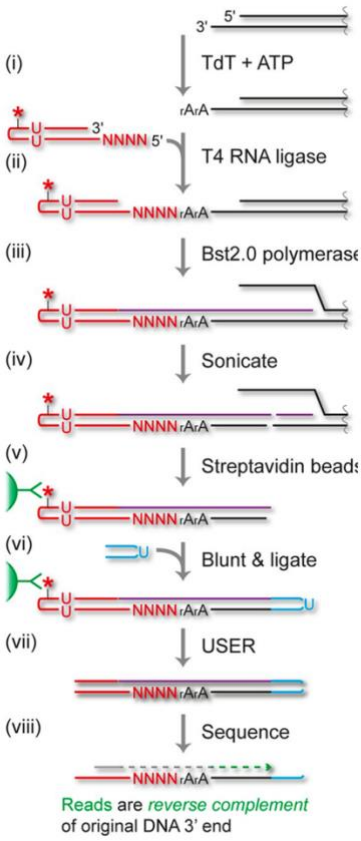

Figure 1-14. TrAEL-seq workflow. Figure adapted from Neesha et al. 2021.

TrAEL-seq not only accurately mapped the DSBs in genome wide, but also detected the stalled replication forks at replication fork barrier sites of yeast and human by targeting the fork reversal events or the potential cleavage of DNA for the replication restart. By calculating the read polarity in asynchronous wild-type cells detected by TrAEL-seq, a significant strand bias is observed and highly corelated to the detected Okazaki fragments distribution indicated that TrAEL-seq can also detect the replication fork progression. Indeed, TrAEL-seq that avoids the pulse-label incorporation and needs fewer cells (less than million human ES cells) to generate a high-quality RFD profile compared to OK-seq, surely provides an encouraging alternative method to study DNA replication and genomic instability in different cell types within various stress conditions. Nevertheless, based on our comparison, performance in yeast of TrAEL-seq is high-resolutive but still need to be improved in human cells (see *Chapter 2-2.3* for more detail).

- **eSPAN**

Recently, some new methods developed meant to reduce the quantity of starting materials and to detect the relevant protein-associated initiation, for instance, checking the relative epigenetic modification during DNA replication. eSPAN (enrichment and sequencing protein-associated nascent DNA) performs stranded sequencing of BrdU or EdU labelled nascent replicated DNA associated with specific histone modifications [10,68]. The original protocol was to ChIP the protein of interest then following by the capture of BrdU-labeled nascent DNA [68]. Then the latest updated version of eSPAN method in mouse cells was using modified CUT&Tag/ACT-seq, a recently developed technique to digest and tag genomic DNA bound by proteins of interest (e.g. H4K20me2, H4K12ac, POLE3/4) using protein A-fused transposase Tn5 (pA-Tn5), followed by enrichment of nascent DNA using BrdU-IP (Fig. 1-15) [10]. Using this eSPAN method, it successfully profiles histone distributions at leading or lagging strands with much fewer starting cells with around 10^6 for mouse embryonic stem cells (mESCs) and even lower

to 50,000 cells for H4K20me2 [10]. Nevertheless, the resolution of mapping is consequently decreased due to the reduction of starting cells.

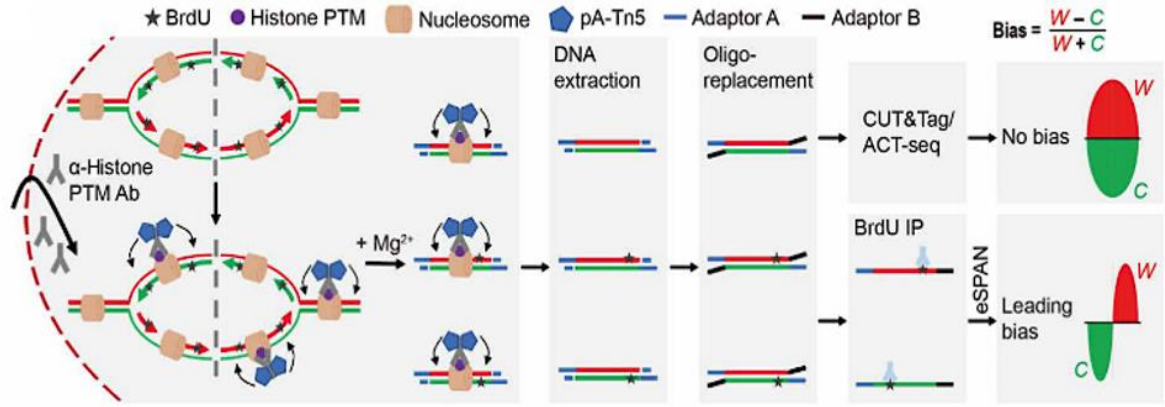

Figure 1-15. Graphic diagram of eSPAN. Figure adapted from Li et al 2020.

- **SCAR-seq**

SCAR-seq (sister chromatids after replication by DNA sequencing) is a method that uses a similar principle to eSPAN. At the latter it also investigates the histone modification during DNA replication by tracking H4K20me2 or H4K5ac [9]. Cells were labeled with EdU, and nascent mononucleosomes carrying H4K20me2 or H4K5ac are immunoprecipitated. EdU is coupled with biotin though click reaction and streptavidin beads are used to recover the newly synthetized DNA. Captured fragment are denaturated to separated parental and the new strand. Strand-specific sequencing is performed and the genome-wide sister chromatid histone proportion can be calculated as: $Partition = (F - R)/(F + R)$ which F is the reads counts of forward strand and R refers to the reverse strand in a defined genomic binsize (Fig. 1-16). MCM2-2A mutants with disruption of histone binding are also generated by gene editing for the comparison with wild type. It revealed that MCM2 mutant defective in histone binding in mouse ES cells also show defects in the transfer of modified parental histones, indicating a conserved role of MCM2 in parental histone transfer [9].

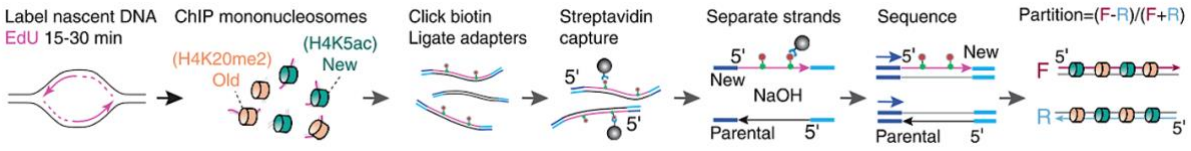

Figure 1-16. Schematic protocol of SCAR-seq. Partition of old and new histones is calculated as the proportion of forward (F) (red) and reverse (R) (blue) counts in genomic windows [the range is between −1 (100% reverse strand) and 1 (100% forward strand)]. Figure adapted from Petryk et al 2018.

2. Coordination between DNA replication and transcription

2.1. Gene transcription

The simultaneous expression of thousands of genes in the nucleus of the eukaryote cells is a strongly controlled process. Transcription is the first step in gene expression, in which genetic information is used to generate a functional product such as a protein. The goal of transcription is to make an RNA copy of a gene's DNA sequence. For a protein-coding gene, the RNA copy, or transcript, carries the information needed to build a polypeptide (protein). Eukaryotic transcripts need to go through some processing steps before translation into proteins. The main enzyme involved in transcription is the RNA polymerase, which uses a single-stranded DNA template to synthesize a complementary strand of RNA. Specifically, RNA polymerase builds an RNA strand in the 5’ to 3’ direction, adding each new nucleotide to the 3’ end of the strand. The three main ones are Pol I, Pol II and Pol III. Each is responsible for the expression of specific transcripts. Pol I is active in the transcription of ribosomal RNAs (rRNA), Pol II of messengers RNA (mRNA) and certain non-coding RNAs (ncRNA), and Pol III of tRNA, rRNA and small non-coding RNAs.

Gene transcription, similarly to DNA replication, also takes place in three stages: initiation, elongation, and termination. To initiate the transcription, RNA pol II binds to a CpG-rich sequence of DNA called the core promoter, 35-40 bp upstream and downstream of the transcription start site (TSS). The most important core promoter element is the TATA box that 5-20% occupies in mammalian and serves as specific binding site for general transcription factors including the pol II and TFIID complex (TBP, TATA-box binding protein) [69]. Other promoter motifs Inr (initiation element, direct initiation even without TATA box), DPE (downstream promoter element) serve for a wide variety of regulatory factors that control the expression of individual genes.

Once recruited on DNA, the RNA polymerase proceeds melting the double helix to provide a single-stranded template needed for transcription. As it "reads" this template one base at a time, the polymerase builds an RNA molecule out of complementary nucleotides, making a chain that grows from 5' to 3'. The RNA transcript carries the same information as the non-template (coding) strand of DNA, but it contains the base uracil (U) instead of thymine (T). Once the RNA transcript is complete by terminators sequences, they release the transcript from the RNA polymerase. In bacteria, RNA transcripts can act directly as messenger RNAs (mRNAs).

However, in eukaryotes, this process is much more complicated. The transcript of a protein-coding gene is called a pre-mRNA and must have their ends modified, by addition of a 5' cap (at the beginning) and 3' poly-A tail (at the end). Furthermore, many eukaryotic genes are interspersed with non-coding sequences called introns. Such sequences are transcribed in the pre-mRNAs but they have to be removed during the RNA maturation. This process is called splicing and can be performed either by a ribonucleoprotein complex called spliceosome or by intron self-catalytic activity. The gene coding parts (called exons) so joined will form a mature mRNA (Fig. 1-17), which is then exported through nuclear pores to cytoplasm were it can be translated by ribosomes [70]. Some genes can be alternatively spliced, leading to the production of different mature mRNA molecules from the same initial transcript [71]. Then RNAP II is paused by recruiting terminators and eventually terminates the transcription with the pre-mRNA cleavage and polyadenylation.

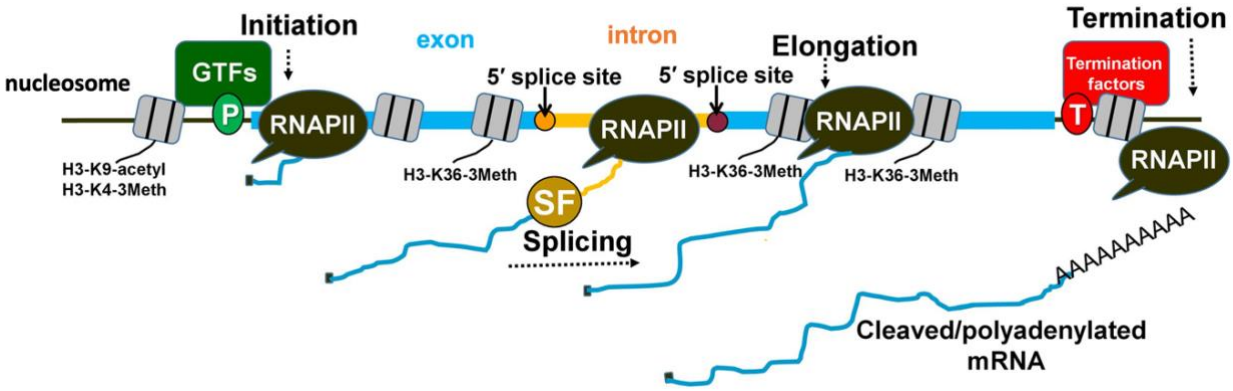

Figure 1-17. Transcription in eucaryotes. Splicing factors are recruited cotranscriptionally to the intron with the help of the RNAPII carboxy-terminal domain. Spliceosomal assembly on the splice sites can facilitate the stabilization of general transcription factors (GTFs) at the promoter region of the gene and prime nucleosomes with activation marks (H3-K9 acetylation and H3-K4 trimethylation) for initiation. The splicing factors can also interact with transcription elongation factors and influence nucleosome modifications (H3-K36 trimethylation) to promote elongation. Similarly, splicing factors can contribute to enhanced termination of transcription by facilitating the recruitment of termination factors and removal of elongation marks that block effective termination. Adapted from Dwyer et al. 2021

An aberrant transcription termination can disrupt coordination of replication and transcription either by affecting origin firing and/or fork progression, or caused by changing the length of the cell cycle phases (shorter G1, longer S phase), resulting in genomic loci being transcribed and replicated simultaneously, leading to potential collision between these two processes and then induce the genome instability [72].

2.2. Conflict between replication and transcription

During replication, forks encounter a variety of protein complexes that are acting onto DNA, such as the transcriptional machinery. In fact, transcriptional complexes are constantly traveling a large part of the genome. On one hand, this has a positive impact on genome stability since allows to quickly detect DNA lesions and to repair it by a mechanism, called DNA repair, coupled with transcription that can occur across the cell cycle, even in S-phase when actively transcribed loci still need to be efficiently duplicated. On the other hand, during the phase S, the transcription complexes represent a hidden threat to replication by causing potential collisions with replisomes, particular in gene-dense regions, which constitutes a major source of endogenous replicative stress leading to genomic instability [63]. Another major risk comes from extremely long genes, whose transcription takes more than one cell cycle, on these loci, the encounter of the replication and transcription machinery is unavoidable. Some of these genes are at common fragile sites (CFSs), loci that replicate late in S phase and are hotspots for chromosomal instability [73]. Transcription-replication conflicts (TRCs) can occur either co-directional (CD) where the replication fork progress in the same direction as transcription or head-on (HO) where the two machineries converge (Fig. 1-18). There are strong evidences revealing that the latter is much more detrimental for genome stability [3,74]. Indeed, replication forks are fragile structures, the blocking of which can induce DNA double-strand breaks (DSB) of DNA and chromosomal rearrangements contributing to the appearance of tumors.

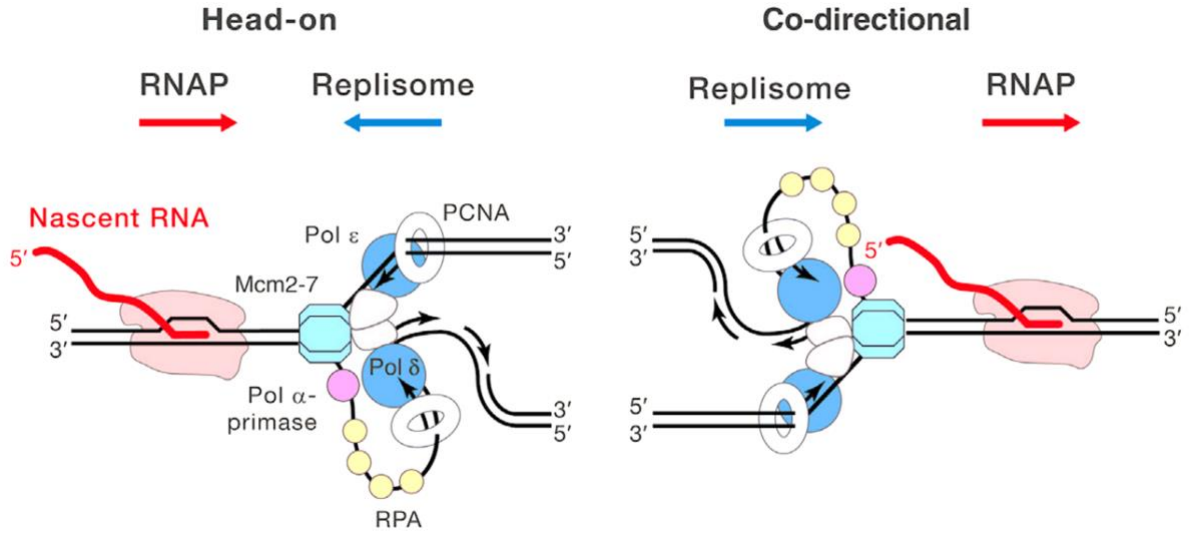

Figure 1-18. Transcription-replication conflicts. The replication-transcription conflict can occur in a co-directional or a head-on manner. Also, the replication fork could encounter the three-stranded nucleic acid structures: R-loops, which could cause the replication fork stalling. Figure adapted from Hamperl et al. 2016

To avoid these conflicts, various mechanisms have been put in place during evolution. In bacteria, which has a circular chromosome and a single replication origin, the transcription of the majority of genes is oriented in a co-directional way with replication [74]. Eukaryotic cells adopt similar mechanism and most genes are replicated codirectionally with transcription [75]. This organization of the genome makes it possible to reduce the frequency of head-on collisions, which are particularly deleterious [76]. Though, in eukaryotes the situation is more complex because the genome is organized in several linear chromosomes, each carrying a large number of replication origins. Not all of these origins are effective and the direction of replication of a given region can vary from one cell to another within cell population [12]. For instance, some long genes that replicate late in S phase can cover with some chromosomal instability loci and some highly transcribed short genes are likely close to early replication origins which are called early replicating fragile sites, which makes conflicts with the inevitable transcription process [3].

3. R-loop

Under endogenous or exogenous stress, triplex structures of RNA:DNA hybrids, that are normally transient, can be stabilized leading to pathological outcomes. These hybrids are called R-loops and it is becoming clear that they are important determinants for genome stability during replication [3].

3.1. R-loop discovery and double-edged functions

R-loops are three-stranded structure preferentially formed in G-rich regions of transcribed genes. They are DNA-RNA hybrid with a displaced non-template DNA strand whose length can span from 200 bp to several kilobases. The first R-loop structure was discovered in 1976 using *in vitro* electron microscopy [77]. In the following years, R-loops were identified in bacteria and other organisms [78]. Since 2003, year in which the Human Genome Project was completed, we have been able to map R-loops on the human genome and infer their functions (Fig. 1-19).

In physiological conditions, R-loops are involved in the immunoglobulin class switch recombination mechanism of stimulated B cells [79]. R-loops are as well enriched at both TSS and TTS of highly transcribed genes where they regulate the nearby epigenetic landscape and gene expression [80]. For instance, on promoters containing CpG island (CGIs), the presentence of an R-loops prevent the deposition of *de novo* methylation keeping these functional elements

active [81]. Moreover, R-loop localized at promoters can serve as recruitment platforms to recruit histone modifiers [4].

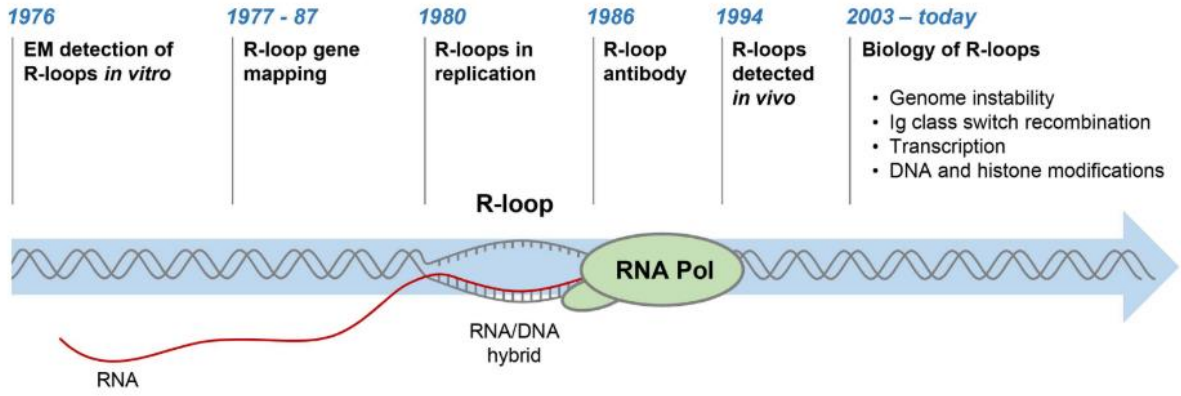

Figure 1-19. The timeline of R-loop research. EM: electron microscopy. Figure adapted from Groh et al. 2014

While R-loops formed on promoters may either stimulate or repress transcription, those formed at the 3' end of gene ensure an efficient and regulated termination of transcription (ref). When RNA Pol II elongates near to the downstream of gene poly(A) regions where are G-rich pause sites, the formation of an R-loops causes the Pol II to pause, then helicase Senataxin (SETX) is recruited to resolve them allowing exonuclease XRN2 to cleave poly(A) sites and so that terminate transcription [82].

R-loops are also detected in mitochondria, where the replication occurs on a circular template in a strand-asynchronous fashion. Here R-loops serve as primer for replicative DNA polymerase and prevent transcription during replication progress [83]. Another beneficial function of R-loops is to promote a faithful chromosome segregation in mitosis. Centromeric R-loop formation can be coated with RPA to activate the ataxia telangiectasia mutated and Rad3-related (ATR) kinase leading to Aurora B activation, which is necessary for accurate chromosome segregation [84].

In spite of these positive functions, it is also evident that R-loops can be a detrimental source of DNA damage, particular in S phase, that can lead to genomic instability [85,86]. This is the case of the R-loop that form at the site of HO TRCs causes DNA damage at collision sites, especially in pathological conditions such as perturbations in BRCA1/BRCA2 pathways or in absence of resolutive factors such as SETX or RNH where the persistent R-loop might become barrier to hinder DNA resection leading to chromosomal rearrangements [87]. Therefore, R-loops need to be tightly regulated in cells to maintain their functional features and to prevent the toxic R-loop accumulation that may lead to human diseases.

3.2. R-loop prevention and degradation

Programmed R-loops facilitate the regulation of gene expression, but their unscheduled presence could become a deleterious source of replication stress if they are not resolved efficiently. In recent years the pathways involved in regulation of R-loop accumulation have been unveiled in eukaryotes (Fig. 1-20):

i) The most efficient enzymes are ribonucleases including all types of RNases H (H1 and H2) that bind DNA-RNA hybrids and degrade the RNA. Of these, RNase H1 is particularly important and its over expression has been shown to repress R-loops induced replication stress [88]. Other players are the endonucleases of the nucleotide excision repair (NER), e.g., XRN2, XPF, XPG, FEN1, that can directly resect the DNA-RNA hybrid;

ii) Helicases in human, such as SETX, Fanconi anemia group M protein (FANCM), Bloom Syndrome RecQ Like Helicase (BLM), DEAD-Box Helicase DDX family (DDX1/5/19/21/23), AQR and so on, can unwind the R-loops in different processes [2,89]

iii) Topoisomerases 1 and 2A are indispensable to suppress the R-loop formation by release the supercoiling stress [87];

iv) Recent studies have found that RNA methyltransferases METTL3 or TRDMT1 can modulate the RNA:DNA hybrids at double strand breaks sites by forming the METTL3-m6A-YTHDC1 axis [90].

v) Based on different biological processes in which R-loops are involved, DNA repair/recombination factors such as BRCA1/2, FANCD2, the chromatin modifiers, transcription termination factors and also the RNA processing factors such as the splicing factor ASF/SRSF1, can also participate the R-loop degradation [89].

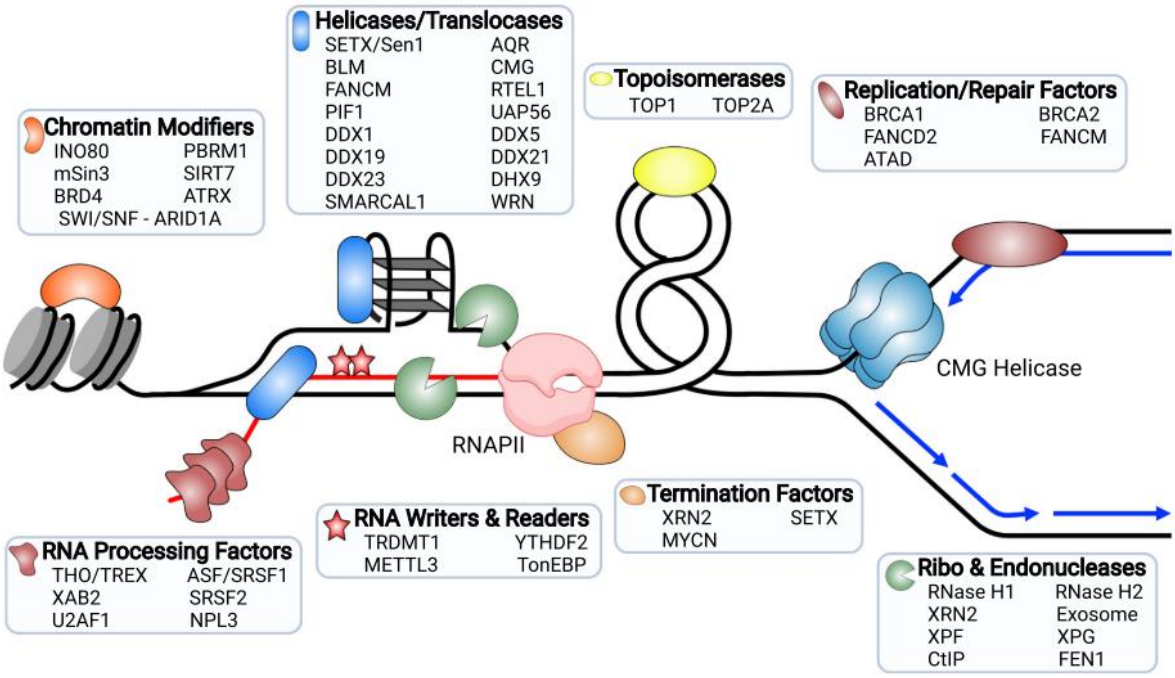

Figure 1-20. Factors that prevent and resolve R-loops. Factors that prevent the formation of R-loops or play a role in their resolution. The black line is parental DNA, blue lines are daughter DNA, and RNA is red lines. The stacked structure on the ssDNA strand of the R-loop represents a G-quadruplex. Figure adapted from Brickner et al. 2022

3.3. R-loop-dependent genomic instability

3.3.1. TRC-induced R-loops, replication fork stalling and restart

More and more approaches to date revealed that, besides all their beneficial features, R-loops may consider as threats to the genomic integrity. The abnormal regulatory factors increase R-loop accumulation. When they cannot be resolved efficiently, it could cause R-loop-dependent genomic instability in eukaryotic cells [87]. Strong evidences suggest that collisions between replication and transcription machinery, especially head-on collision, might stabilize the formation of R-loops and causing the replication fork to stall and possibly leading to DNA damage [3]. As mentioned in the previous session, TRC can either occur in co-direction (CD) or head-on (HO) manner. Cimprich and colleagues found an increase of R-loops colocalized with stalled replication forks in HO collisions while a decrease level of R-loops is observed in CD collisions indicating that different types of TRCs may lead to distinct regulating pathways [3].

During DNA replication, when the forks encounter transcription machinery in an HO way, the polymerases might have to pause the progression either within a direct converging conflict or an increased positive supercoiling stress between the two machineries. Stable R-loops are formed and the increasing torsional stress may fold the G-rich sequences into G-quadruplex

(G4) structures, which can also induce the fork stalling [89,91]. However, if the two machineries meet co-directionally, the replication forks also can be stalled when the RNA polymerase backtrack towards replication or again by the R-loops [89].

To avoid causing the further DNA damage, forks need to restart efficiently. For the CD TRC situation, the restart of the fork requires the CMG complex to either bypass or unwind the hybrid. In the first case, the DNA Pol α will be recruited to restart the progression; in the second case the hybrid is removed by the CMG complex and DNA replication continues (Fig. 1-21A, B). The situation is more complicated if the RNA polymerases is bound to the R-loop. In this case the passage of the CMG complex could unwind the R-loop freeing the RNA Pol that would proceed transcription in front of the replicative fork, or alternatively, the RNA Pol has to be removed or the CMG complex must bypass it. (Fig. 1-21C, D).

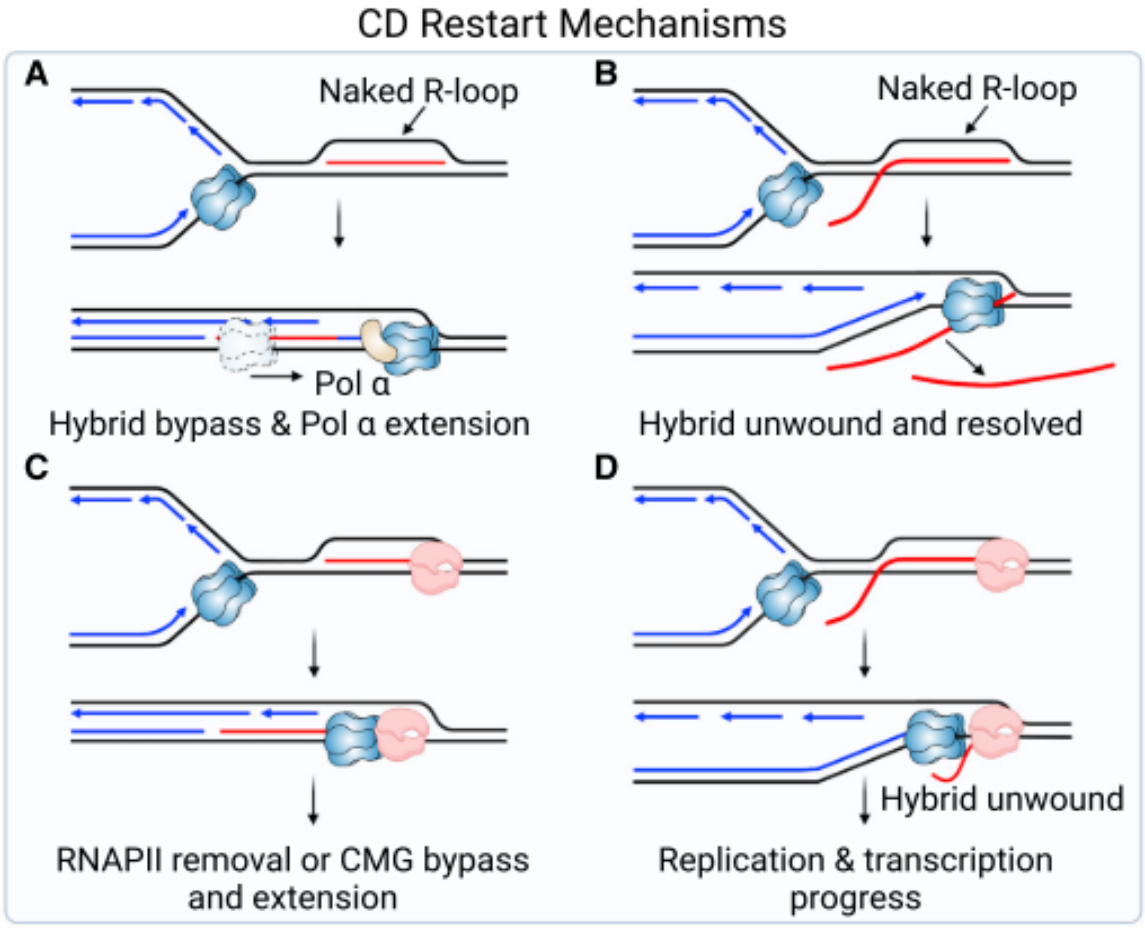

Figure 1-21. Co-directional restart of replication stalling mechanism. (A) The CMG helicase translocates over and bypasses a ‘‘naked’’ hybrid with RNA annealing to DNA and RNA strand can be extended by Pol α. (B) The CMG helicase unwinds a hybrid when the 5’ end of the RNA strand is exposed as a flap, allowing fork progression to continue. (C) CMG translocation over the hybrid would lead to its arrest at RNAP. Fork progression require CMG to bypass the stalled RNAP or for RNAP to be removed. (D) CMG unwinding of a hybrid bound to RNAP allow RNAP to continue and replication fork to progress behind transcription. Figure adapted from Brickner et al. 2022

The HO restart mechanism is more complicated. If a replication fork only encounters a naked R-loop, it can restart in the same way as CD. Though, when the fork encounters RNA pol, different pathways can take place. If the CMG helicase can bypass the RNAP, PrimPol which has both primase and polymerase activities is employed to restart replication. When the replication completed, RAD18 and UBE2B can be recruited to promote the gap filling during post-replicative repair [92] (Fig. 1-22A). This, therefore, could potentially induce hybrid formation through the gap existence by ssDNA forming behind the fork, which provide a pathway for *de novo* RNA synthesis or post-replicative hybrid formation by recruiting again RNAP (Fig. 1-22B). The third potential restart mechanism involves an intermediate replication reversal process. In this context, nuclease SLX4, MUS81 and its partner protein EME1 play important role to cleave the excessive forks, resolve the R-loops. The recombinase RAD51-mediated replication fork reversal may occur to stabilize the stalled fork. If it does, the helicases RECQ1 is required to reset the reversed fork while RECQ5 removes RAD51 at the R-loop stalled fork. To conclude, replication is restart by RAD52 and POLD3 following the elongation factor ELL mediated transcriptional restart and LIG4/XRCC4-mediated fork re-ligation [89] (Fig. 1-22C).

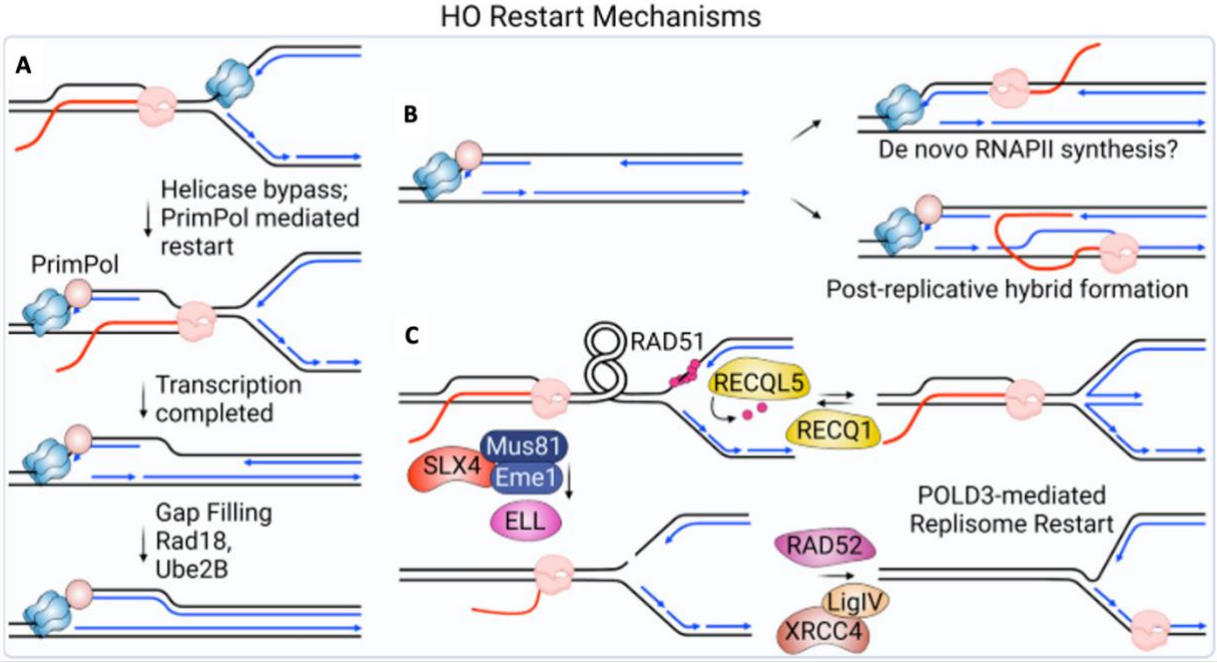

Figure 1-22. Head-on restart of replication stalling mechanisms. (A) CMG bypass of RNAP would leave ssDNA exposed, allowing PrimPol recruitment. Primer synthesis by PrimPol allows the fork to continue, leaving behind ssDNA that can be replicated by a gap-filling mechanism once transcription is complete. (B) ssDNA resulting from bypass and repriming or a failure to reprime, and which persists behind the replication fork, could serve as a template for the formation of hybrids. Hybrids could form through de novo synthesis by RNAP (top) or through the association of nascent RNA resulting from transcription after the fork has passed (bottom). (C) Restart of a stalled fork initiated by SLX4 and MUS81-mediated fork cleavage is thought to relieve the torsional stress, resulting in resolution of the R-loop formed during a HO conflict. Religation of the fork by XRCC4/LIG4 and ELL-mediated restart of transcription allows both replication and transcription to continue. Prior to fork cleavage, RAD51 may promote fork reversal. RECQ1 is needed to reset the fork and RECQ5 promotes removal of RAD51. Figure adapted from Brickner et al. 2022.

3.3.2. DNA damage response to the R-loop-related TRCs

Once the R-loops cannot be resolved in time or due to the inefficient regulatory factors under replicative stress, DNA damage could be sequentially produced. Cells therefore need to react to such damage activating a DNA damage response (DDR) to repair it faithfully. DNA damage can occur either on both strands (double-strand breaks, DSB) or on only one strand (single-strand breaks, SSB). In the first case, the damage is sensed by the MRN complex (MRE11-RAD50-NBS1) that recruits Ataxia-telangiectasia mutated (ATM) on the damage and activates its pathway. SSB are instead sensed by the accumulation of RPA on ssDNA that causes the recruitment of ATM and RAD3-related (ATR) through ATRIP allowing the activation of its pathway [93,94]. In CD-TRC, an ATM pathway activation can be observed due the production of DSBs. This can happen through different mechanism: (i) Damage on the displaced ssDNA of an R-loop can promote the formation of DSB at the replication fork passage; (ii) nucleases recruited to resolve the R-loops can cut on both strands; (iii) or the RNAP backtracks towards the replication fork (Fig. 1-23A). Meanwhile in HO collisions, ATR pathway is initiated by SSBs with an accumulation of RPA-coated ssDNA at the stalled forks (Fig. 1-23B) [3,87].

The DDR can also regulate R-loop resolution pathways, for instance, by promoting the recruitment of SETX to the collision site or causing the translocation of DDX from nuclear pore in nucleus to unwind the hybrid by ATR activation [87]. However, due to the double-edged functions of R-loops, how to distinguish the only deleterious ones in the whole genome and the precise mechanism about theses R-loop-dependent TRC impact the genomic instability and how activate the different DDR pathways are still unclear.

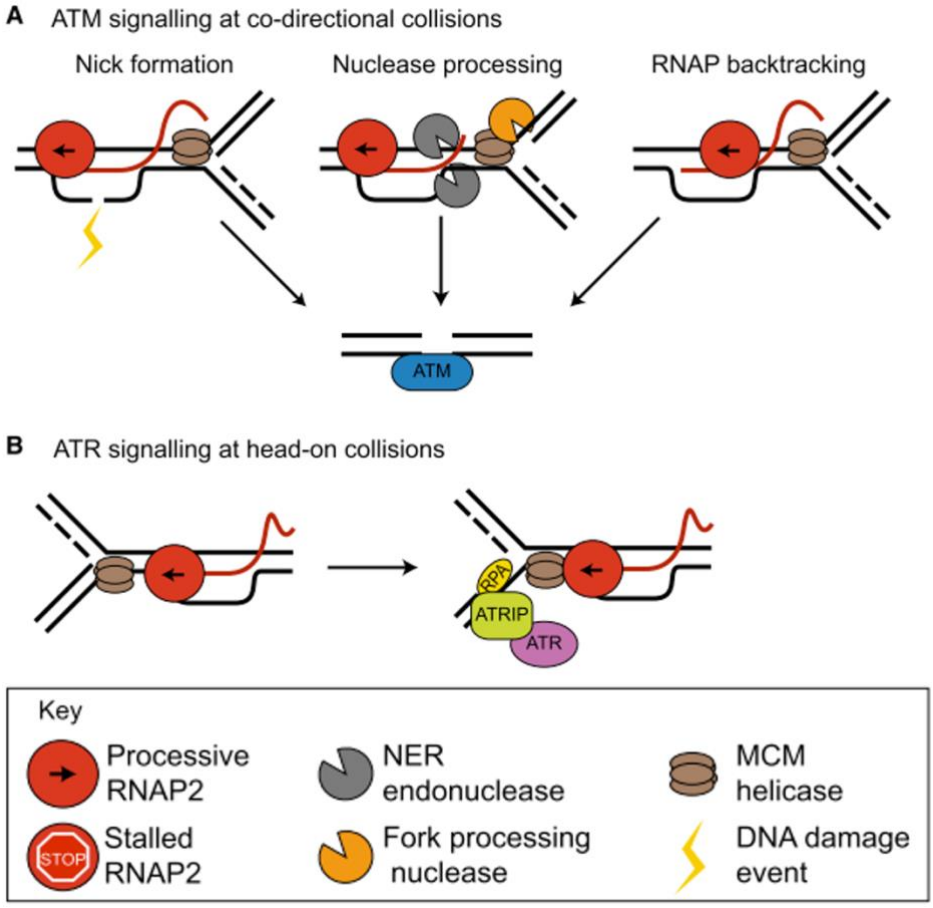

Figure 1-23. DNA damage response to R-loops. (A) R-loops may cause DSBs and ATM activation through three mechanisms: damage to the displaced ssDNA that is converted to a DSB by DNA replication, nucleolytic processing of the R-loop or fork stalled by R-loops, or collision of the replication fork with a backtracked RNA polymerase. (B) R-loops may activate ATR in head-on collisions by stalling replication forks with accumulation of RPA. Figure adapted from Crossley et al. 2019

4. Objectives of my Ph.D.

The faithful transmission of genetic information to daughter cells is central to maintaining genome stability. In human, at each cell division, tens of thousands of replication origins need to be activated to ensure complete duplication of >6 billion nucleotides of the genome. The replication program is routinely exposed to endogenous and exogenous stresses, which plays an important role in many human diseases. In particular, these alterations can represent an important early cause of cancer. Strong evidence in recent years supports that oncogene-induced replication stress is a major driver of tumor progression. During last decade, gene transcription has been discovered, by us and the others, as a major source of endogenous stress that can affect the replication forks progression by causing TRCs, leading to fork stalling, DNA collapse, and genome instability. Recent evidence revealed that TRCs can be regulated by R-

loops but due to the limitation of TRC research both in biology and in informatics analysis, the concrete mechanism of R-loop-associated TRC to regulate the genome stability is still unknown.

The objective of my PhD therefore is to try to fill this gap. In the following *Chapter 2* session, I will precisely describe the bio-informatics toolkit named OKseqHMM that I developed for analyzing the OK-seq related data, which is the first bioinformatics tool available to date to analyse the RFD data obtained from various techniques with an accurate prediction for the genome-wide replication initiation zones, termination zones and even the flanking replication transition regions based on hidden Markov model (HMM). This tool can be applied widely and compatibly from yeast to human cells and user-friendly for the usage, which facilitates the scientific community in the same research field to deeper investigate the replication-related projects. **Please noted, the context of *Chapter 2* corresponds to my bioinformatics method paper is available on *bioRxiv* (1st author)** [95] and is accepted to publish in *Nucleic Acid Research*. In the meanwhile, collaborated with Olivier Hyrien's lab members, the complete protocol with experiment description of OK-seq together our tool usage paper is accepted to publish in *Nature Protocol* (co-1st author, Annex 1). In tightly collaboration with Philippe Pasero's lab (Institute of Human Genetics, Montpellier) who is expert in R-loop and TRC research, I will present in *Chapter 3* about our recent results about TOP1 as a guardian to prevent TRC-associated R-loops formation and to maintain the genomic stability in human cancer cells. Based on the results of analysing multi-omics data, we successfully built our model how R-loops impact the replication progression in normal and TOP1-deficient cells. **The results and methods shown in *Chapter 3* are already published in a *Nature communications* paper of which I am the co-1st author** [96] **and a related *author's view* paper is published in *Molecular and Cellular Oncology*** [97] **(1st author) (Annex 2)**. **The figures which generated by our collaborators/colleagues will be indicated in the corresponding figure legends.** In the *Chapter 4*, I will mainly discuss the perspectives that our study can be further applied and its indispensable role to investigate the replication-related domains. In addition, a scientific review about replication and transcription written together with colleagues in our lab was published in 2020 (2nd author) [2] (Annex 3).

Chapter 2 OKseqHMM

During each cell division, tens of thousands of DNA replication origins are co-ordinately activated to ensure the complete duplication of the entire human genome. However, the progression of replication forks can be challenged by numerous factors. One such factor is transcription-replication conflicts (TRCs), which can either be co-directional or head-on with the latter being revealed as more dangerous for genome integrity. In order to study the direction of replication fork movement and TRCs, we developed a bioinformatics toolkit called OKseqHMM (https://github.com/CL-CHEN-Lab/OK-Seq) and used it to analyse a large number of datasets obtained by Okazaki fragment sequencing (OK-seq) in organisms including yeast, mouse and human, to directly measure the genome-wide replication fork directionality (RFD) as well as accurately predict the replication initiation and termination at a fine resolution. We further successfully applied our analysis to extensive related techniques, which also contain RFD information (e.g., eSPAN, TrAEL-seq). Our works, therefore, provide an important tool and resource for the community to further study TRCs and genome instability, in a wide range of cell line models and growth conditions, which is of prime importance for human health.

The progression of replication forks can be challenged by numerous factors. One such factor is transcription-replication conflicts (TRCs) since the replication and transcription machineries share the same DNA template. TRC can either be co-directional (CD) or head-on (HO), and the latter has been revealed as more detrimental for genome integrity [3]. Previous bioinformatics analyses have revealed that, in large numbers of genomes from bacteria [98] to human [63,99], most genes are co-directionally oriented with replication forks to avoid the more deleterious HO TRC. Recently, a new method to directly measure the genome-wide replication fork directionality (RFD) along the human genome by sequencing of Okazaki fragments (OK-Seq) [12], which are present only on the lagging replicating strand, allows quantitatively analysing and accurately detecting replication initiation and termination. The analysis of OK-seq data of human cells has also demonstrated a significant co-direction of replication fork progression with gene transcription within active genes [12].

More and more techniques are now being developed, for instance, Pu-seq [6], eSPAN [10], SCAR-seq [9], GLOE-seq [8] and TrAEL-seq [11], which also provide genome-wide RFD information. Moreover, in recent years, strong evidence shows that replication- and transcription-related mutational asymmetries are widespread across cancer development [100]. Especially (Apolipoprotein B mRNA editing enzyme, catalytic polypeptide) APOBEC-associated

mutations (also called APOBEC mutation signatures) in humans are represented in up to 15% of all sequenced tumours and contribute to 50% of all mutations in many tumours. APOBEC-associated mutations preferentially occur on the lagging-strand template during DNA replication, and are also highly associated with mismatch repair and transcription-coupled damage repair in cancer [101–105]. Furthermore, N6-methyladenosine (m6A) modifications have been considered as one of the most prevalent internal modifications in mammalian mRNAs and the abnormal m6A modification caused by m6A modulators, e.g., methyltransferase-like 3 (METTL3), is a common feature of various tumours [106–108]. Evidence has shown that METTL3 and m6A could promote homologous recombination-mediated repair of double-strand breaks (DSBs) by modulating RNA:DNA Hybrid (R-loop) accumulation [90]. Importantly, R-loops have been recently shown, by others and us, to be frequently accumulated at transcription termination sites of actively transcribed genes displaying high HO TRCs [96,97]. Therefore, systematically unveiling the genome-wide DNA replication panorama is essential for human health.

Despite its importance, to date, there is no published available tool to analyse RFD data and to determine the replication initiation and termination positions genome-wide, although several methods have been previously described for OK-seq data analysis, such as using the Hidden Markov Model (HMM) to analyse human OK-seq data [109] or the origin efficiency metric (OEM) to analyse yeast OK-seq data [54,110]. It is, therefore, important to have a uniform framework of OK-seq data (and related data) analyses. Here, we developed a bioinformatics toolkit, called OKseqHMM, to directly obtain the high-resolution RFD profile genome wide. Besides the fork direction, the toolkit also deciphers the information of replication initiation/termination zones using an algorithm based on HMM, calculates the OEM to visualize the transition of RFD profile at multiple scales, and finally generates the average metagene profiles and heatmaps to provide RFD/OEM distributions along the regions of interest (Fig. 2-1). We have gathered a large number of published available OK-seq data (13 in total) from *S. cerevisiae*, mouse and human cells, and successfully obtained the high-resolutive (~1 kb for mouse and human cells and ~50 bp for yeast) RFD profiles and the accurate calling of corresponding replication initiation and termination zones genome-wide.

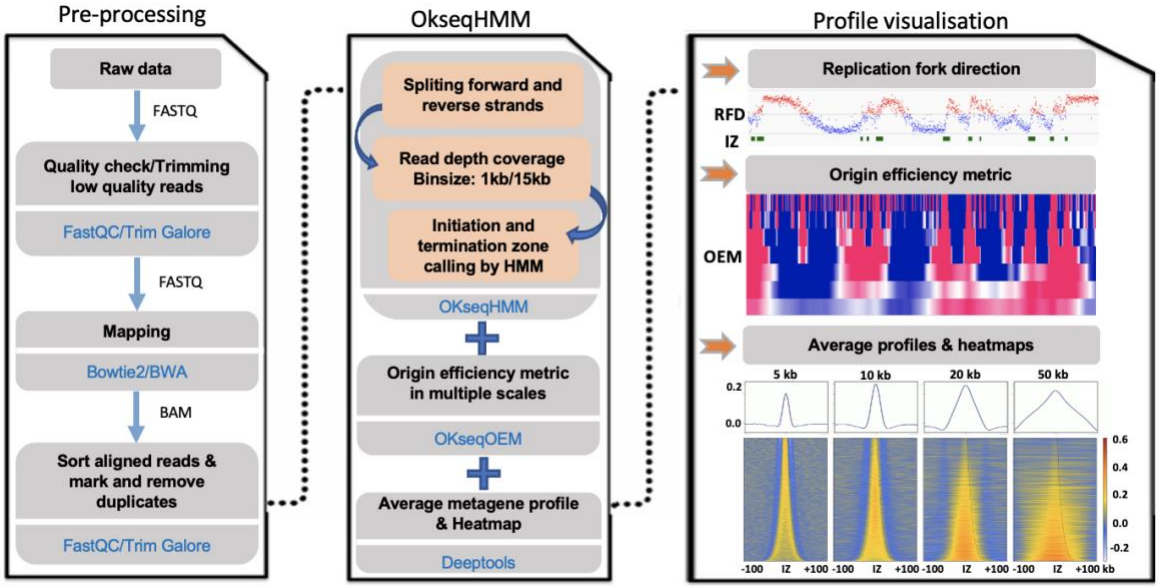

Figure 2-1. Schematic presentation of data analysis pipeline of OKseqHMM toolkit. The raw sequencing data can be pre-processed into aligned files by corresponding bioinformatics tools indicated in blue (left panel). The middle panel shows the major functions of the OKseqHMM toolkit. The first function of OKseqHMM checks the input aligned bam files to determine whether they are single- or paired-end sequencing data, then automatically splits the reads into Watson and Crick strands and calculates the replication fork directionality (RFD). By default, the calculation is performed within 1 kb adjacent windows (recommended for human cells) and then smoothed into 15 kb sliding windows with 1 kb step. These parameters can be easily adjusted based on the nature of the data. Different replication features, i.e., initiation zones (IZ), two intermediate states and termination zones, are predicted based on an HMM algorithm (See Implementation for detail). The second function (OKseqOEM) uses the reads on Watson and Crick strands to generate origin efficiency metrics (OEMs) at multiple scales to visualize the RFD transition. And the last function allows users to generate an average metagene profile and heatmap to analyse distributions of RFD and OEM around the genes/regions of interest. Results can be visualized in the genomic visualization browsers (such as IGV) as shown in the right panel.

1. Material and methods

OKseqHMM toolkit is an R package for profiling OK-seq data to study the genome-wide replication program. This R package contains multi-functions and is served for analysing OK-seq data from the original mapping bam file(s) to count matrices, RFD calculation, initiation/termination zone calling and average metagene profiles/heatmaps.

1.1. HeLa S3 OK-seq data generation and sequencing

HeLa S3 cells were cultured in DMEM high glucose medium with 10% FBS. Ok-seq Libraries were generated starting from exponentially growing cells as previously described [109,111]. Libraries were sequenced on an Illumina NextSeq system using PE (75 cycles). The other cell lines are indicated in the table.1.

1.2. Function OKseqHMM measures RFD and predicts replication initiation/termination zones

OKseqHMM is the main function of the toolkit, which involves several important steps of OK-seq data analysis. The function transforms OK-seq data into RFD profiles for a primary visualization (e.g., with the genomic visualization browsers, such as IGV), then, it can accurately identify replication initiation zones (IZs, upward transitions on RFD profile), termination zones (TZs, downward transitions on RFD profile) and also the intermediate states (flat RFD profile) along the genome by using the HMM.

For each window, RFD was computed as follows:

$$\boldsymbol{RFD} = \frac{\boldsymbol{C} - \boldsymbol{W}}{\boldsymbol{C} + \boldsymbol{W}}$$

where $\boldsymbol{C}$ and $\boldsymbol{W}$ correspond to the number of reads mapped on the Crick and Watson strands, which reveal, respectively, the proportions of rightward- and leftward- moving forks within each window (e.g., 1 kb window was used for OK-seq data of human cells). Since the total amount of replication on both strands should be constant across the genome, we normalized the difference between the two strands by the total read count to account for the variations in read-depth due to copy number, sequence bias and so on. RFD ranges from -1 (100% leftward-moving forks) to +1 (100% rightward-moving forks), and 0 means equal proportions of leftward- and rightward-moving forks. Data obtained from biological replicates produced RFD profiles that strongly correlated to each other, for HeLa cells, Pearson R=0.92, $p<10^{-15}$ (t-test) and for GM06990 cells, R=0.93, $p<10^{-15}$.

A four-state HMM was used in OKseqHMM to detect within the RFD profiles the ascending (AS), descending (DS) and flat (FS) segments representing regions of predominant initiation ('Up' state), predominant termination ('Down' state) and constant RFD ('Flat1' and 'Flat2' states) [109] (Fig. 2-2A). In the HMM segmentation process, the RFD values were computed within 15 kb (for human OK-seq data) sliding windows (by default, stepped by 1 kb across the autosomes). The HMM used the ΔRFD values between adjacent windows, in which $\Delta RFD_n = \frac{RFD_{n+1} - RFD_n}{2}$ for the window *n*. By default, windows with <30 reads on both strands were masked. The ΔRFD values (also between -1 and 1) were divided into five quantiles and the HMM package of R (http://www.r-project.org/) was used to perform the HMM prediction with probabilities of transition and emission, which are manually defined by the training dataset (Fig. 2-2B). The same transition and emission probabilities used in our previous study [109] were set

as default values and used in all OK-seq data analyses in the current study. Two representative examples of human RFD profiles together with the segments of IZs, TZs and two Flat states obtained by OKseqHMM were shown in Fig. 2-2C, D. The choice of a 15 kb sliding window is based on a compromise between spatial resolution and reproducibility of AS detection among biological replicates. Finally, the efficiency of the detected AS (i.e., initiation zones) was estimated as:

$$\Delta RFD_{segment} = \frac{RFD_{end} - RFD_{start}}{2}$$

where RFD_{start} and RFD_{end} correspond, respectively, to the RFD values computed in 5 kb windows around the left and right extremities of each segment.

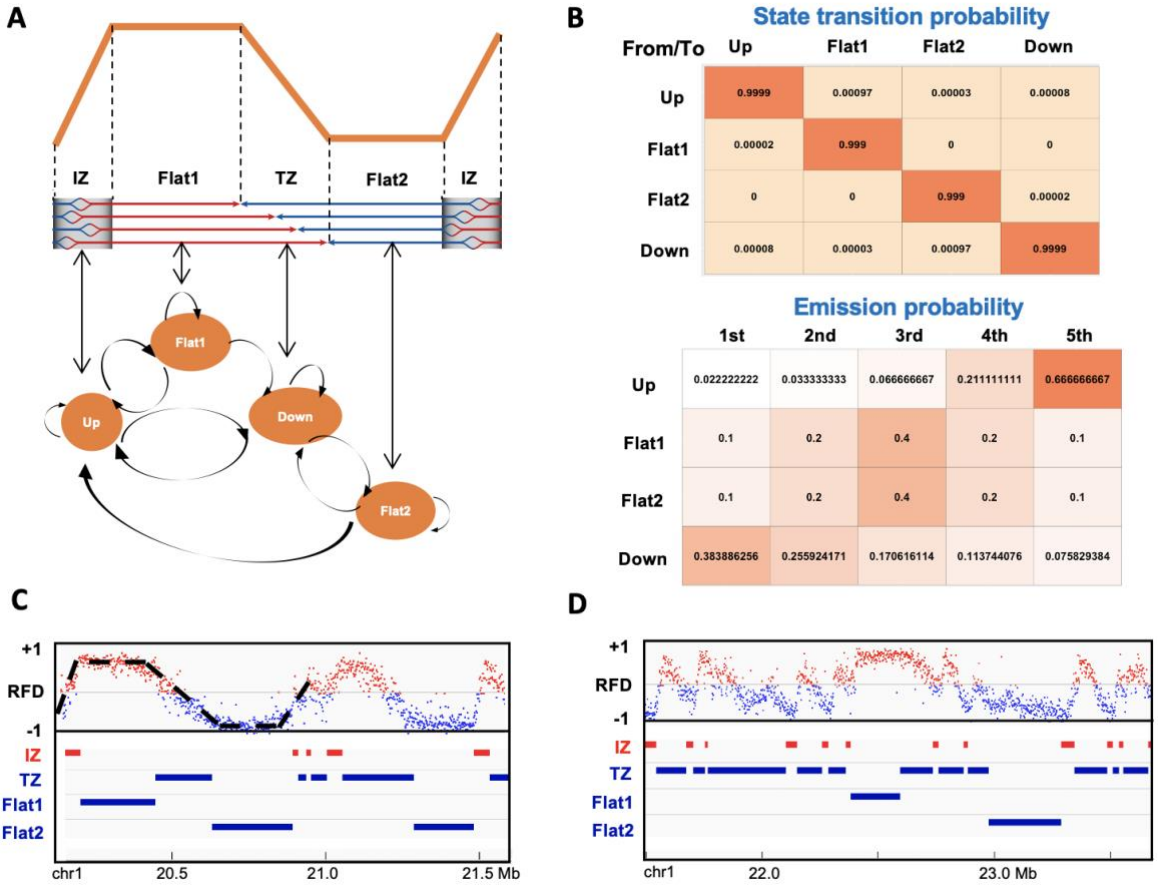

Figure 2-2. Schematic presentation of HMM algorithm for initiation and termination zone detection. (A) A 4-state HMM model used in the segmentation process: Up, regions of predominant initiation (IZ); Down, regions of predominant termination (TZ); Flat1 and Flat2, two intermediate transition states. (B) Default state transition probability (between states) and emission probabilities (probabilities of each state within five quantiles of ΔRFD values) used in OKseqHMM (see Material and Methods for detail). The probability matrixes were colour coded based on their values (higher probability values are closer to red). (C) and (D) Examples of RFD profile of human HeLa cells from chromosome 1 with the corresponding IZs, TZs and 2 Flat states identified by OKseqHMM. Each point on the RFD profile gives the RFD value calculated within each 1 kb adjacent window, and the windows with positive and negative RFD values are shown in red and blue, respectively.

1.3. Function OKseqOEM generates the RFD transition profiles at multi-scales

For a further investigation of origin efficiency (i.e., ΔRFD), we provide here a second function to visualize it directly at multiple scales. As defined in the previous publication for yeast OK-seq data analysis [23], the density of Okazaki fragments on the Watson and Crick strands are compared within 4 fixed-size sliding bins, which are strand-specific 10 kb quadrant values to calculate an Origin Efficiency Metric (OEM), computed as $OEM = \frac{W_L}{W_L+C_L} - \frac{W_R}{W_R+C_R}$ (W_L and W_R measure, respectively, the read density in the left and right quadrants on the Watson strand, while C_L and C_R refer to the density on the Crick strand), ranging from -1 to 1 for each base in the genome. Maximal values in the OEM scores represent replication origins, while the minimal ones are considered as regions of replication termination. In addition, the different amplitudes of positive OEMs (from 0 to 1) are referred to as origin-firing efficiency; and the degree of termination at each position can be measured from 0 to -1.

Here, we further extend OEM calculation within a fixed window size into multiple-scales to better fit OK-seq data analysis of other organisms, such as human cells.

$$\boldsymbol{OEM_{i\ for\ list[n]} = \frac{\left(W_{i+list[n]} - W_i\right)}{\left(W_{i+list[n]} - W_i\right) + \left(C_{i+list[n]} - C_i\right)}}$$

Where ***list[n]*** can be defined by users as a list of windows (e.g. [1,10, 20, 50, 100]), ***i*** is from 1 to the total length of the data – *list[n]*. ***C*** and ***W*** correspond to the number of reads mapped on, respectively, the Crick and Watson strands within corresponding windows.

Using the two bam files of reads within, respectively, Watson and Crick strands generated by the previous OKseqHMM function and the annotation coordinates, the function OKseqOEM can automatically calculate the OEM profiles at a series of defined scales (e.g., from 1 kb to 1 Mb for human cells), which allows us to directly visualize the transition states of replication and to validate the IZs identified by OKseqHMM then to double-check the size and boundary of IZs.

1.4. Average metagene profile/heatmap provides RFD distribution on specific genomic regions

To analyse RFD distributions around and/or among the genomic regions of interest, such as the identified IZs or the annotated genes, we developed an additional module for the metadata

analysis. With the gene coordinates (or IZs) together with the RFD and/or OEM big wiggle files generated from OKseqHMM and/or OKseqOEM functions, we can easily obtain the corresponding profiles/heatmaps by using the computeMatrix and plotProfile/plotHeatmap functions of deepTools (https://deeptools.readthedocs.io/en/develop/index.html) [112] via defining the genomic distances of interest for the upstream and downstream borders.

2. Results

2.1. Genome-wide RFD and replication origin detection in yeast

To evaluate the performance, we first applied our tool to the available yeast OK-seq data [7]. OKseqHMM was successfully applied to the yeast OK-seq data to generate the RFD profile at a fine resolution (50 bp), the OEM profiles at different scales (from 50 bp to 25 kb) and a precise IZ/Origin calling (Fig. 2-3A). About 350 robust IZs were finally identified by OKseqHMM, which range from 0.5 kb to 5.5 kb with an average length of 1.5 kb (Fig. 2-3B, Table 1). To check the accuracy of IZ calling results, we compared OK-seq IZs with the known yeast origins, i.e. autonomously replicating sequence (ARS) from OriDB 2.1.0 [113], and up to 70% of our detected IZs were found at ≤ 2 kb distance (between centres) from a known ARS. As expected, the OK-seq IZs are better correlated with the confirmed ones, with 185, 36, and 22 overlapped (i.e., distance between centres ≤ 2 kb) with the confirmed (median distance 0.27 kb), likely (median distance 0.48 kb) and dubious (median distance 0.69 kb) origins, respectively (Fig. 2-3C). In case we consider all OriDB origins instead of only the overlapped ones, the distances between OK-seq IZ centres to the closest OriDB origins of each class are still significantly smaller (median distance 0.41, 1.13 and 1.77 kb for confirmed, likely and dubious origins, respectively) than random simulated genomic positions (Fig. 2-3D).

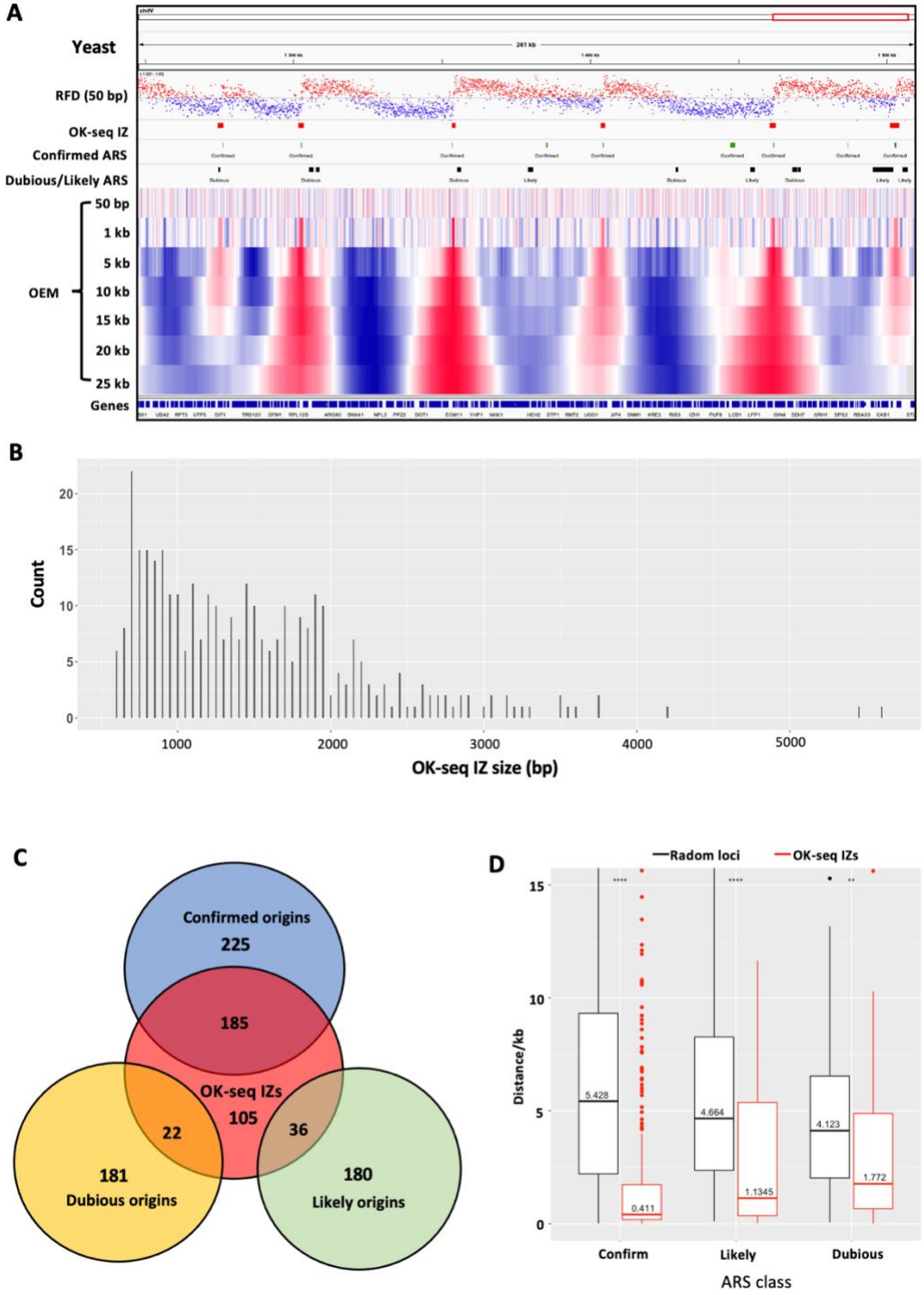

Figure 2-3. Analysis of OK-seq data by OKseqHMM in Yeast. (A) Yeast RFD profile was calculated at 50 bp resolution with the corresponding IZs identified by OKseqHMM, which are highly correlated with the confirmed origins from OriDB (29). RFD profile as in Fig. 2C, but with 50 bp resolution. Below, the OEM profiles calculated from 50 bp to 25 kb scales, and the windows with positive and negative OEM values are shown in red and blue, respectively. (B) Length distribution of detected OK-seq IZs. (C) Venn diagram showing the overlap numbers of OK-seq IZs compared with all the known origins clustered in 3 categories (confirm, likely and dubious) from OriDB, in which overlap means that the closest distance between each other's centres is less than 2 kb. Note that all confirmed OriDB origins are not overlapped with an OK-seq IZ since all origins might not active in the culture condition and/or yeast strain used for the OK-seq experiment. Further comparison with origins identified by other techniques can be found in Fig. 7B. (D) The boxplot shows the distribution (in red) of distances between an IZ centre detected by OkseqHMM and the closest origin centre from OriDB for 3 categories, which is much closer compared with the random simulation control (in black) indicated with significance levels of Wilcoxon rank sum test; ** < 10-2, *** < 10-3, **** < 10-4.

2.2. Genome-wide RFD and replication initiation zone detection for different human cell lines

We then further applied the OKseqHMM to analyse the OK-seq data of human cells. In addition to the published OK-seq data of HeLa MRL2 cells [109], we also generated new additional OK-seq data from HeLa S3 cells, a widely used Encode Tier 2 cell line. The RFD profiles of the two HeLa cell lines are very similar ($R=0.86$, $p<10^{-15}$), suggesting that a similar replication program and IZ positions are used (Fig. 2-4A). The correlation between two HeLa cells is slightly lower than the correlation between two biological replicates ($R=0.92$, $p<10^{-15}$) of HeLa MRL2 cells [109], suggesting that the differences between the two HeLa cell lines represent true biological differences and not only technical variation. About 10,000 IZs have been identified in each HeLa cell line (Table 1), two-thirds of which are common between the two cell lines (Fig. 2-4B). The conservation of IZs is even higher in the early-replicating regions, with 80% of early IZs being shared between the two HeLa cell lines (Fig. 2-4B). A very striking difference of human RFD data compared to those of yeast is that, instead of a sharp 1 kb upward transition of RFD at fixed yeast origins, the size of upward transition of RFD, therefore the IZ length, of human cells is around 10-50 kb (average ~30 kb, ~20-folds larger than the IZ of yeast) (Fig. 2-4A, Table 1). The heatmaps of OEM profiles computed around IZs at different scales show the strongest positive signals at the corresponding scales, i.e., 10 kb scale for the small IZs (<10 kb), 20 kb scale for the IZs of mid-size (20-50 kb) and 50 or 100 kb for the large IZs (>50 kb), respectively (Fig. 2-4C), confirming that RFD transition is associated with the detected IZ length. This further supports the difference between the yeast and human OK-seq pattern and the accuracy of IZ detection obtained by OKseqHMM.

Replication initiation has been previously reported to be enriched within intergenic regions between active genes [109]. To demonstrate how our toolkit can help in the analysis of the association between DNA replication and gene transcription, we analysed the average profiles and the corresponding heatmap of expression level (RNA-seq and GRO-seq) for all detected IZs sorted by their length and confirm that gene transcription presents immediately surround the IZs while with a much lower level within IZs (Fig. 2-5A). To further compare the distribution of RFD and gene transcription, we calculated the average RFD profile and the corresponding heatmap around TSSs (transcription start sites) and TTSs (transcription termination site) of 16,336 active genes (RPKM > 1) in HeLa cells with an extension ± 50 kb upstream or downstream (Fig. 2-5B). This clearly indicates a frequent replication initiation (upward transition of RFD) at both regions upstream of TSS and downstream of TTS, which leads to a

co-direction between replication and transcription at TSS while a higher head-on TRC at TTS, in agreement with previous publications [96,109].

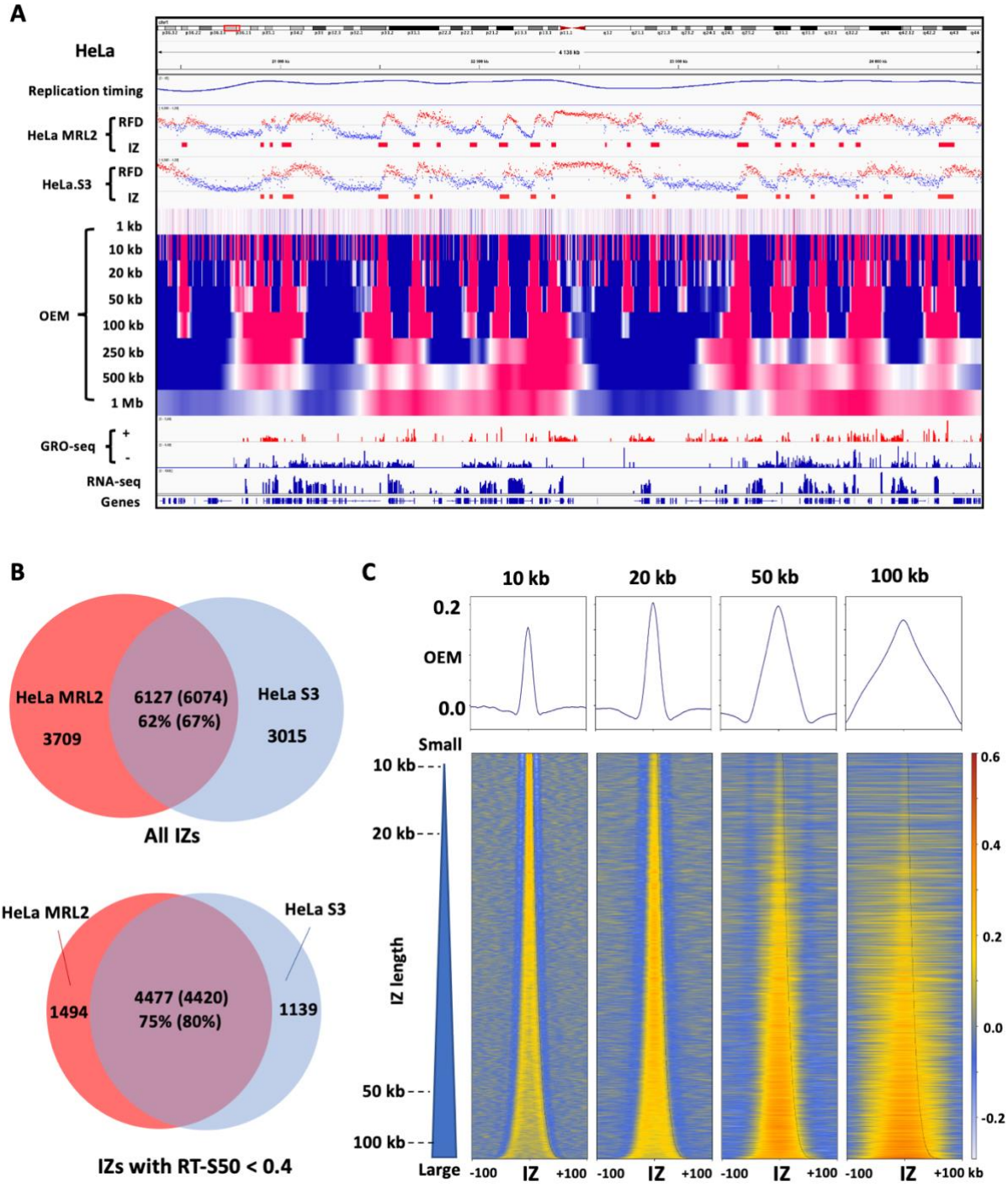

Figure 2-4. Analysis of OK-seq data of HeLa cells by OKseqHMM. (A) Replication timing profile obtained by Repli-seq, RFD profiles and corresponding detected IZs for published HeLa MRL2 OK-seq data (6) and OK-seq data of HeLa S3 cells generated in the current study, the OEM profiles of HeLa S3 cells from 1 kb to 1 Mb scales, and the transcription data provided by GRO-seq and RNA-seq along a ~4 Mb region on chromosome 1. (B) Venn diagrams showing that two-third of Ok-seq IZs matched between the two HeLa cell lines and the overlap goes up to 80 % for the early IZs (with replication timing S50 < 0.4). (C) Mean OEM profiles and heatmaps of OEM (heatmap colour scale is indicated on the right) around the HeLa S3 IZ centers at indicated scales (i.e., 10, 20, 50 and 100 kb) sorted by the length of detected IZs.

In addition to OK-seq data of HeLa cell lines, we have gathered and reanalysed with OKseqHMM, the OK-seq data from previous publications for a large amount of human cell lines of different cell types [109,111], such as fibroblast (IMR90), lymphoblastoid (GM06990) and lymphoma (Raji, BL79, IARC385), leiomyosarcoma and leukaemia (IB118, TLSE19, K562), and erythroblast (TF1) (Table 1, Fig. 2-6). OKseqHMM generated high-quality cell-type-specific RFD profiles and robust IZ calling for all data analysed. The sizes of IZs in different cell types are within the same range (average size between 26 to 36 kb), demonstrating that it is a common feature of human cells. Nevertheless, we observed that the RFD profiles of different cell lines are quite conservative for some origin-rich regions while they are cell-type specific. The data obtained with the cell lines of close cell type or origin show similar pattern of RFD profiles (Fig. 2-6B), for instance, Pearson correlation R is up to 0.87 between Raji and BL79 cells since both are lymphoma cells, and close to GM06990 human lymphoblastoid cells, with a Pearson R=0.79.

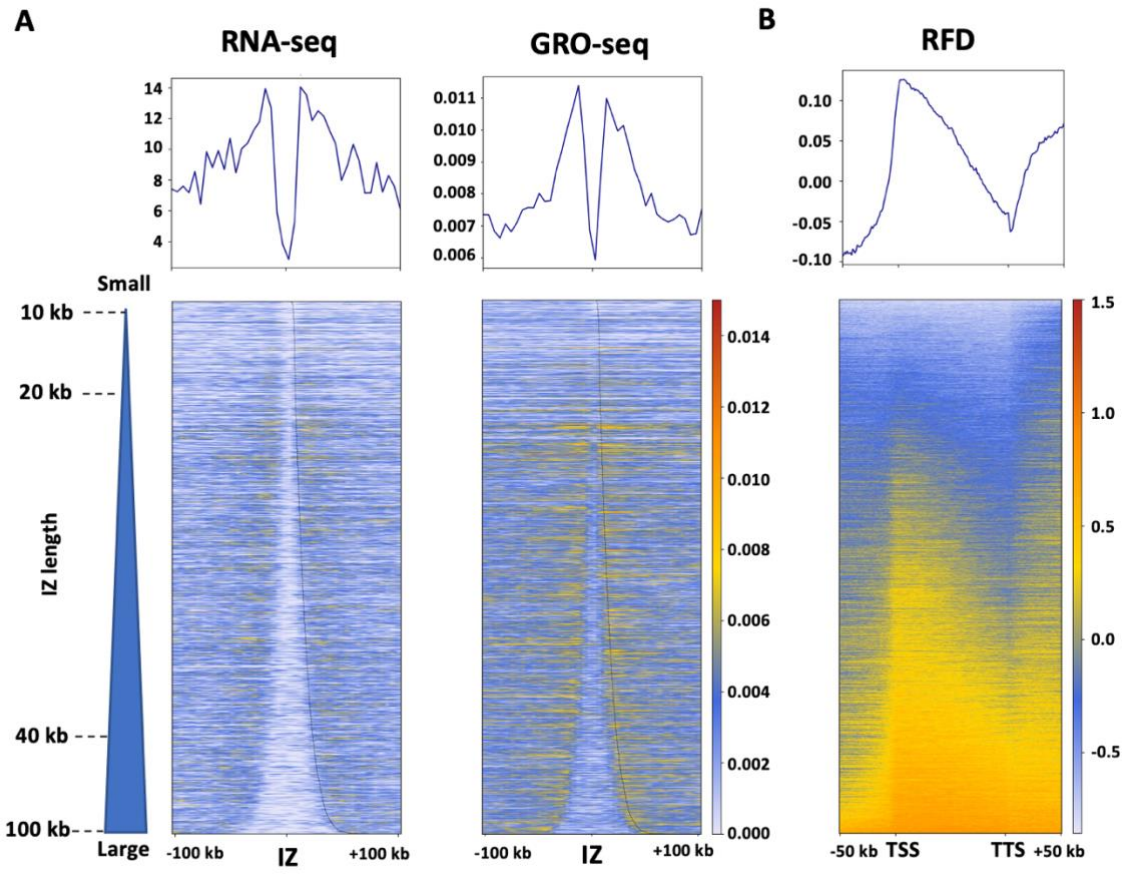

Figure 2-5. OKseqHMM reveals the coordination between DNA replication and gene transcription. (A) Mean profiles and heatmaps of RNA-seq and GRO-seq around the HeLa S3 OK-seq IZ centres. (B) Mean profile and heatmap of HeLa S3 RFD between TSS (transcription start site) and TTS (transcription termination site) of active genes with an extension of +/- 50 kb. The heatmap colour scales are indicated in each panel.

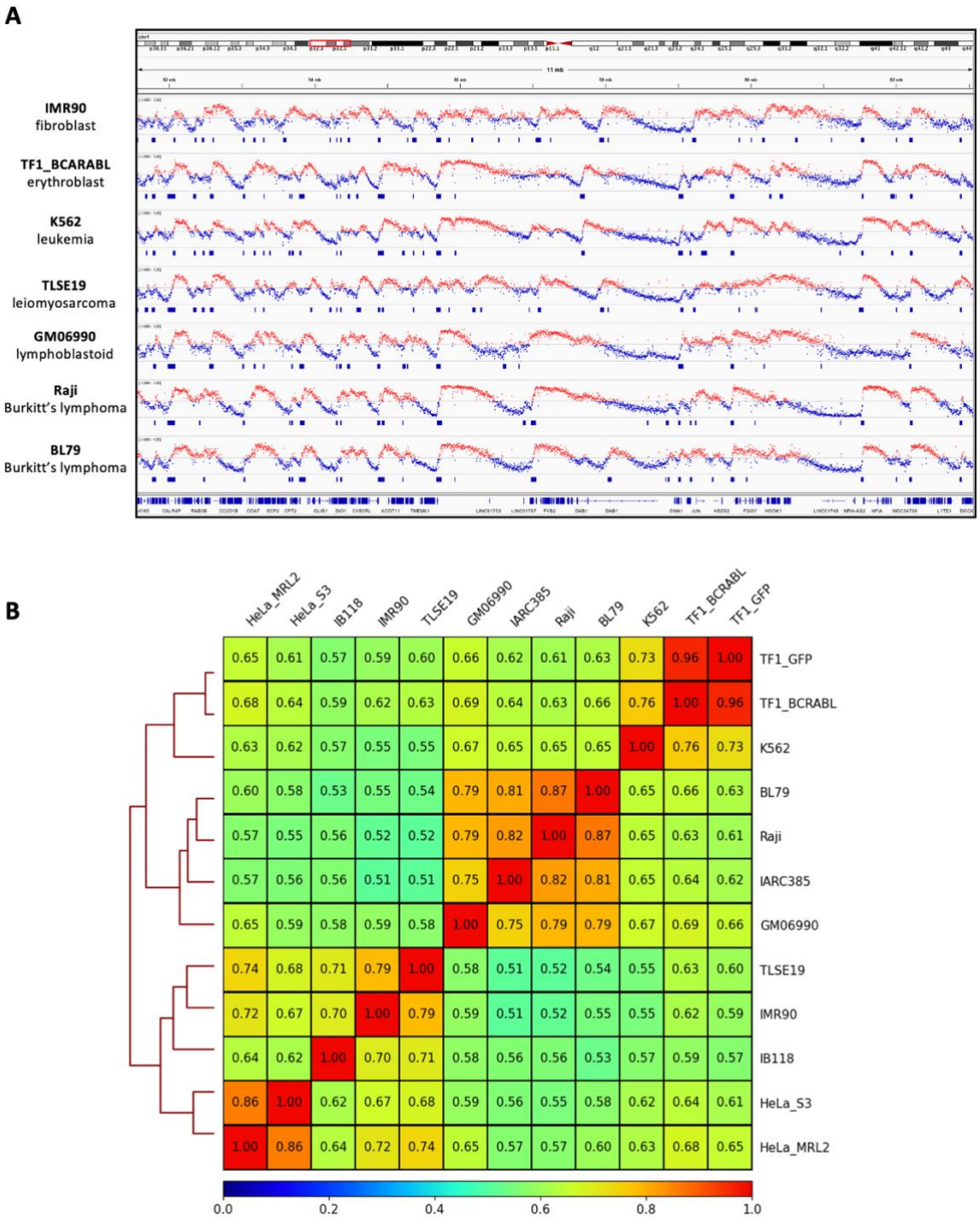

Figure 2-6. Genome-wide RFD profiles of different human cell lines show the cell-type-specific replication program. (A) Cell-type-specific RFD profiles and the corresponding detected IZs for indicated human cell lines, IMR90, TF1, K562, TLSE19, GM06990, Raji and BL79. (B) Pairwise Pearson correlations between OK-seq data (1 kb) of different human cell lines.

2.3. Extend OKseqHMM to analyse the RFD profiles from other sequencing data

In addition to OK-seq, the OKseqHMM toolkit can be applied to calculate RFD profiles from the sequencing data obtained with other related techniques. As a demonstration, we further extended our toolkit to analyse the published eSPAN [10] and TrAEL-seq [11] data. The RFD data computed from the yeast TrAEL-seq data are very close to those obtained by OK-seq (Fig. 2-7A, R=0.93, $p<10^{-15}$), and the RFD profile obtained by TrAEL-seq even shows a higher quality with less local noise than the OK-seq RFD profile. This difference does not seem to result from the fact that the TrAEL-seq data used in the analysis have higher coverage (contain about two-fold more reads) compared with the available OK-seq data, because TrAEL-seq data always show a less local noise profile after down-sampling to the same coverage as OK-seq.

To further evaluate the IZs detected in different techniques, we here also integrate the IZs identified with FORK-seq [7], a nanopore sequencing based method that allows mapping replication initiation within single DNA molecules. The comparison between the TrAEL-seq IZs, OK-seq IZs, FORK-seq and yeast ARSs showed that 70% (243/348) of OK-seq IZs and up to 84% (321/380) of detected IZs from TrAEL-seq were found within 2kb distance from a known ARS. 79% (191/243) of OK-seq IZs associated with ARSs were also detected by TrAEL-seq and around 77% (186/243) of them were found in FORK-seq (Fig. 2-7B). Notably, a small percentage of initiation sites that are not associated with OriDB origins are robustly detected by OK-seq, TrAEL-seq and FORK-seq, supporting the previous observation that replication initiation in yeast can also occur, although with low frequency, at loci barely enriched in ACS (ARS consensus sequence) motifs [7].

Finally, we also successfully applied OKseqHMM to the OK-seq and eSPAN data [10] of mouse embryonic stem cells (mESC). However, due to the lower amount of reads for the available dataset of eSPAN data, although we used a larger window size (e.g., 10 kb smoothing window instead of 1 kb window) we still got too noisy RFD profiles to perform a robust IZ calling, which reflected that our detection method is read depth dependent. Even so, we still obtained a mean RFD profile similar to those of OK-seq around the IZs identified in the mESC OK-seq data (Fig. 2-7C).

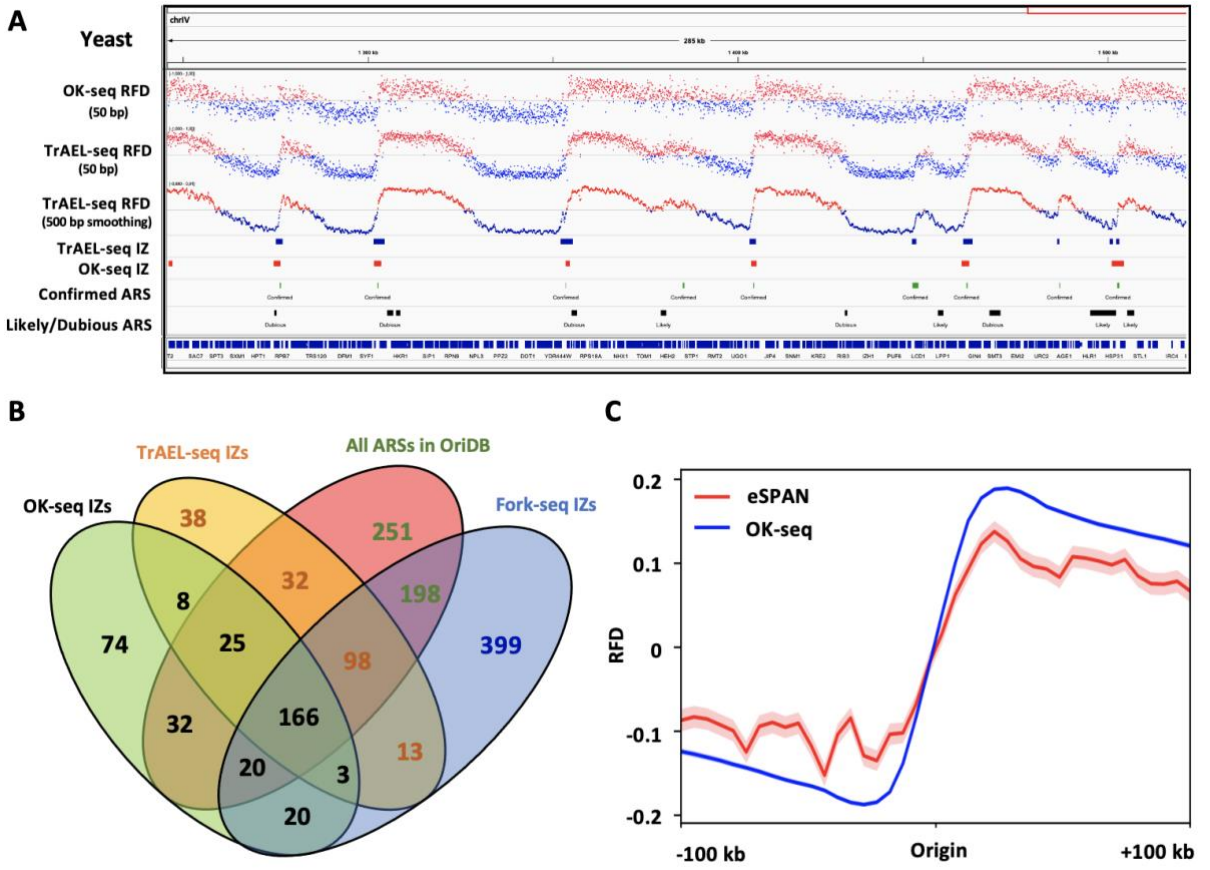

Figure 2-7. Genome-wide RFD profiles obtained from TrAEL-seq and eSPAN data. (A) RFD profiles and the corresponding IZs in 50 bp bin size of OK-seq and TrAEL-seq data of yeast (11). The known origins (ARSs) are downloaded from OriDB. (B) Venn diagram showing the overlap between OK-seq IZs (n=348), TrAEL-seq IZs (n=380), FORK-seq initiation events (n=4964) and published origins (ARSs) from OriDB (n=829), in which overlap means that the closest distance between each other's centre is less than 2 kb. In case origins of one dataset overlap with several origins of other datasets, only one number was provided with the following priority order: OK-seq > TrAEL-seq > OriDB > FORK-seq. It should be noted that there are more origins unique for FORK-seq, since it's a single-molecule technique that allows identification of initiation events with very low frequency. (C) Metagene average RFD profiles computed from OK-seq of mouse embryonic stem cells (mESC) and H4K20me2 eSPAN data of MCM2-2A mutant cells (8). Mean and standard error bands are shown for both data, while the standard error bands of OK-seq data are too narrow to be seen.

3. Discussion

Genome-wide replication fork directionality data have become an important key in understanding numerous biological processes, such as transcription-replication conflicts, replication-associated mutagenesis, replication couple epigenetic maintenance, etc. Here, we present OKseqHMM, a comprehensive R package, to analyse OK-seq data from various cell types and species to generate and visualize high-resolutive RFD and OEM profiles along the genomes, as well as generate the average profiles/heatmaps on the regions/genes of interest. The toolkit also allows accurate detection of replication initiation/termination zones with an HMM algorithm. To our knowledge, this is the first bioinformatics tool available to date to

handle and analyse the RFD data obtained from various techniques. The toolkit is based on R, which should be easily used for both bioinformatics as well as biologists.

We successfully applied OKseqHMM to a large amount of available OK-seq data from different species, including yeast, mouse cells as well as numerous normal and cancer human cell lines (Table 1). This provides an important resource for large research communities, who are interested in studying DNA replication programs, transcription-replication conflicts, replication-associated chromatin organization, replication-associated mutations, genome instability and cancer genomics, among others. Importantly, in addition to OK-seq, more and more new techniques have been developed to study DNA replication and are also able to provide the replication fork direction information. These include the methods like eSPAN and SCAR-seq performing stranded sequencing of BrdU or EdU labelled nascent replicated DNA associated with specific histone modifications, or like TrAEL-seq and GLOE-seq based on the single-stranded end presented on specific replicative templates. Here, we demonstrated that OKseqHMM can be applied to analyse data obtained by both kinds of techniques, i.e., eSPAN and TrAEL-seq, and obtained high-quality results (Fig. 2-7). Notably, techniques like TrAEL-seq, which do not need to incorporate labels and need fewer cells to generate a high-quality RFD profile compared to OK-seq, will provide a good alternative to study DNA replication and genomic instability in different cell types within various stress conditions.

It should be noted that the initiation parameters, such as the transition and emission probabilities, are defined based on the OK-seq datasets of human cells. Although we have shown in the current study that they are quite robust and can be also applied to OK-seq data of yeast and mouse cells to obtain satisfactory results, they might need to be adjusted based on the sequencing-depth and data quality of other datasets, to have an optimal IZ/TZ calling. In the future, with technical improvement, we might be able to further extend the OKseqHMM to study the extrinsic (cell-to-cell) or intrinsic (homolog-to-homolog) variability of DNA replication, if we can further extend the relative techniques to obtain data at the single-cell level and/or in an allele-specific manner as recently achieved for the replication timing study [40,114].

Name	Cell type	Replicate	Initiation zones		Termination zones		Accession number (Reference)
			Number	Size (kb) Mean ± SD	Number	Size (kb) Mean ± SD	
BL79	Burkitt's lymphoma		7798	29±18	7791	211±244	ENA: PRJEB25180 [111]
GM06990	Lymphoblastoid cells	2*	5684	33±19	5715	182±166	SRA: SRP065949 [109]
HeLa MRL2	Epithelial cell of adenocarcinoma	2*	9836	31±18	9441	141±144	SRA: SRP065949 [109]
HeLa S3	Epithelial cell of adenocarcinoma		9089	32±19	9084	223±245	GEO: GSE193547 (Current study)
IARC385	B lymphocytes from Burkitt's lymphoma		4465	36±19	4455	125±164	ENA: PRJEB25180 [111]
IB118	Leiomyosarcoma		3645	26±16	3640	428±440	ENA: PRJEB25180 [111]
IMR90	Fibroblast		12482	26±17	12468	151±147	ENA: PRJEB25180 [111]
K562	Late-stage chronic myeloid leukemia		6982	28±15	6967	136±158	ENA: PRJEB25180 [111]
mESC E14	Mouse embryonic stem cells		3370	27±14	3347	483±554	GEO: GSE142996 [10]
Raji	Burkitt's lymphoma		8096	29±16	8080	143±135	ENA: PRJEB25180 [111]
TF1	BCR-ABL negative cell line from erythroblast		8377	27±17	8371	196±193	ENA: PRJEB25180 [111]
TLSE19	Leiomyosarcoma		10500	27±17	10492	146±144	ENA: PRJEB25180 [111]
Yeast	*S. cerevisiae*	2*	348	1.5±0.7	787	14±13	ENA: PRJEB36782 [7]

Table 1. All OK-seq data analysed by OKseqHMM. * If data of biological replicates are available, the profiles obtained with the combined data are used in the figures, and only the segments (i.e., IZs and TZs) reproducibly identified in both biological replicates were retained.

Chapter 3 TRC-associated R-loop and genome stability regulated by TOP1

Replication stress (RS) including a variety of endogenous and exogenous events interfered with replication fork progression is a major trigger of genomic instability that has been implicated in cancer development. Following with a bidirectional replication mechanism, Replication forks stall when they encounter obstacles such as secondary DNA structures, highly transcribed genes or tightly bound protein complexes [1]. Prolonged fork arrest may lead to fork collapse and to gross chromosomal rearrangements [115]. Stalled replication forks are detected by the intra-S phase checkpoint, a surveillance pathway sensing the presence of excess ssDNA at damaged forks. This checkpoint response is initiated with the binding of the ATR kinase and its partner ATRIP to the ssDNA-binding protein RPA [94]. Once activated by TopBP1, ATR phosphorylates multiple targets, including the RPA32 subunit of the RPA complex on S33 (called thereafter p-RPA). Fork collapse also leads to the phosphorylation of the histone variant H2AX by ATR (γ-H2AX). Unlike p-RPA, the γ-H2AX signal can spread over several hundreds of kilobases around broken forks [1]. ATR also activates the CHK1 kinase to amplify the checkpoint response, repress late replication origins and prevent premature entry into mitosis [116].

As mentioned in *Chapter 1* session, Transcription-replication conflicts (TRCs) represent a major source of replication stress in all organisms, from bacteria to human [3,5] in head-on (HO) or co-directional (CD) manner and frontal collisions are considered much more deleterious to the genomic stability [5]. Besides the HO collisions, replication forks can also encounter three-stranded nucleic acids structure called R-loops, which have been proposed to play both positive and negative roles in gene expression and other chromosome functions, and normally they can be prevented/removed by Topoisomerase I (TOP1), helicases and endonucleases like RNase H [4]. TOP1, as an enzyme that relaxes DNA supercoiling and prevents R-loop formation, which is considered as a safeguard to maintain the genomic stability. Depletion of TOP1 certainly increases the RS and leads to a R-loop-enriched transcriptional-related damage [117]. In order to unveil the mechanism by which R-loops interfere with replication fork progression and provoke potentially genomic instability in human cells especially under replicative stress, we are collaborating with Philippe Pasero's lab (IGH, Montpelier) to perform the necessary experiments in HeLa wild-type and HeLa TOP1-deficient cells.

We have mapped RNA:DNA hybrids, replication stress markers and DNA double strand breaks (DSBs) in wild-type and cells depleted TOP1 and R-loops were observed at both transcription start sites (TSS) and termination sites (TTS) of highly expressed genes. In contrast, the phosphorylation of RPA signals activated by ATR referred to as stalled replication forks were only detected at TTS regions where are preferentially replicated in a head-on orientation relative to the direction of transcription. In TOP1-depleted cells, DSBs also cumulated at TTS, leading to persistent checkpoint activation, spreading of γ-H2AX on chromatin and leading a global replication fork slowdown. These data indicate that fork pausing at the TTS of highly expressed genes containing R-loops prevents head-on conflicts between replication and transcription and maintains genome integrity in a TOP1-dependent manner [96] .

1. Materials and methods

1.1. Cell culture

HeLa cells (around 5 x 10^6) are used for investigation. TOP1-deficient cells are treated by expressing short hairpin RNAs (shRNA) against TOP1 (i.e., shTOP1 cells). Same experiments performed for shSRSF1 cells with expressing shRNA against SRSF1. HeLa, HEK293T and AsiSI–ER-U2OS cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal calf serum (FCS) and 100 U ml−1 penicillin/streptomycin at 37 °C in 5% CO2.

Production of lentiviral vectors and cell transduction. HIV-1-derived lentiviral vectors were produced in HEK293T cells. To this end, cells were seeded on poly- D-lysine coated plates and transfected with packaging plasmid (psPAX2, Addgene plasmid #12260): transfer vector (pLVX-Tet-on; TRIPZ-shTop1): vesicular stomatitis virus envelop plasmid (pMD2.G, plasmid #12259) at a ratio 5:3:2 by the calcium phosphate method. The culture medium was collected 48 h post-transfection, filtrated using 0.45-μm filters and concentrated at 100 folds by ultra-centrifugation at 89,000 × g at 4 °C for 1.5 h. HeLa cells were transduced at a MOI = 10 (multiplicity of infection) by centrifugation at 1500 × g at 30 °C for 90 min in the presence of 5 μgml^{-1} of Polybrene.

1.2. Related high-throughput sequencing techniques

DNA fiber spreading

To perform DNA fiber spreading, HeLa control and shTop1 HeLa cells were treated with 2 μgml^{-1} doxycycline for 24 h and then transfected with the plasmid EGFP-N1 or RNase H1-EGFP for 48 h in the presence of doxycycline. Subconfluent cells were sequentially labeled first with 10 μM 5-iodo-2′-deoxyuridine (IdU) and then with 100 μM 5-chloro-2′-deoxyuridine (CldU) for the indicated times. One thousand cells were loaded onto a glass slide (StarFrost) and lysed with spreading buffer (200 mM TrisHCl pH 7.5, 50mM EDTA, 0.5% SDS) by gently stirring with a pipette tip. The slides were tilted slightly and the surface tension of the drops was disrupted with a pipette tip. The drops were allowed to run down the slides slowly, then air dried, fixed in methanol/acetic acid 3:1 for 10 min, and allowed to dry. Glass slides were processed for immunostaining with mouse anti-BrdU to detect IdU, rat anti-BrdU to detect CldU, mouse anti-ssDNA antibodies, and corresponding secondary antibodies conjugated to various Alexa Fluor dyes. Nascent DNA fibers were visualized using immunofluorescence microscopy (Leica DM6000 or Zeiss ApoTome). The acquired DNA fiber images were analyzed by using MetaMorph Microscopy Automation and Image Analysis Software (Molecular Devices) and statistical analysis was performed with GraphPad Prism (GraphPad Software). The length of at least 150 CldU tracks were measured per sample.

Detection of pRPA32-S4/S8 foci by immunofluorescence

Cells growing on coverslips were incubated for 3 min at room temperature with CSK buffer (10 mM PIPES, pH 7.0; 100 mM NaCl; 3 mM MgCl2; 300 mM sucrose and 0.3 mg ml^{-1} RNase A) containing 0.7 % Triton X-100 and phosphatase inhibitor cocktail and fixed with 2 % PFA for 10 min at room temperature. The coverslips were incubated with an anti-pRPA32-S4/S8 antibody overnight at 4 °C and then with a secondary antibody conjugated to an Alexa Fluor dye for 1 h at 37 °C, followed by DAPI staining and ProlongGold mounting. Images were acquired by using a Zeiss LSM780 confocal or a Zeiss ApoTome microscope. The mean fluorescence intensity (MFI) in cells was quantified by using CellProfiler (www.cellprofiler.org).

Detection of RNA:DNA hybrids by slot blotting

Cells were lysed in 0.5% SDS/ TE, pH 8.0 containing Proteinase K overnight at 37 °C. Total DNA was isolated with phenol/chloroform/isoamylalcohol extraction followed by standard ethanol precipitation and quantified using Nanodrop. Half microgram of total DNA was loaded in duplicate onto a Hybond-N^{+} membrane using slot blot apparatus. The membrane was separated in two, one for direct UV crosslinking at 0.12 Joules and the other for DNA

denaturation. To denature DNA, membrane was incubated with denaturation buffer (0.5 M NaOH; 1.5 M NaCl) for 10 min and neutralization buffer (1 M NaCl and 0.5M Tris, pH 7.5) for another 10 min prior to UV crosslinking. Membranes were blocked with 5% skim milk in PBST (PBS; 0.1% Tween- 20) for 1 h. The RNA:DNA hybrids and ssDNA were detected by immunoblotting.

Chromatin fractionation

Cells were incubated with CSK-Triton lysis buffer (10 mM PIPES, pH6.8; 100 mM NaCl; 1 mM $MgCl_2$; 1 mM EGTA; 300 mM Sucrose; 10 mM DTT; 0.2% Triton X-100; protease inhibitor; phosphatase inhibitor) on ice for 10 min and harvested by scraping. The supernatant was collected after centrifugation at 0.8 × *g* for 5 min at 4 °C. Pellet was resuspended in CSK-Triton buffer and incubated for 10 min on ice. Another round of centrifugation at 0.8 × *g* for 5 min at 4 °C was performed to separate nucleoplasm and chromatin fractions, supernatant and pellet, respectively.

DNA/RNA immunoprecipitation sequencing (DRIP-seq) [81]

cells (5×10^6) were lysed in 0.5% SDS/TE, pH 8.0 containing Proteinase K overnight at 37 °C. Total DNA was isolated with phenol/chloroform/isoamylalcohol extraction followed by standard ethanol precipitation. One-third of total DNA was fragmented by a cocktail of restriction enzymes (EcoRI, HindIII, BsrgI, SspI, XbaI) overnight at 37 °C. A negative control treated overnight with RNase H was included. Digested DNA was purified by phenol/chloroform/isoamylalcohol extraction, ethanol precipitation and quantified by Nanodrop. Four micrograms of digested DNA were diluted in binding buffer (10 mM NaPO4, pH 7.0; 0.14 M NaCl; 0.05% Triton X-100) and incubated with 10 μg of S9.6 antibody overnight at 4 °C on a rotator. DNA/antibody complexes were added for 2 h at 4 °C to Agarose Protein-A/G beads prewashed with binding buffer. Immunoprecipitated DNA was eluted by incubating with elution buffer (50 mM Tris pH 8.0; 10 mM EDTA; 0.5% SDS) containing Proteinase K at 55 °C for 45 min on a rotator. The eluent was precipitated by phenol/chloroform/isoamylalcohol extraction and ethanol precipitation. Validation of DRIP procedure was performed by qPCR. The pulled down material and input DNA were then sonicated, size-selected, and ligated to Illumina barcoded adaptors, using TruSeq ChIP Sample Preparation Kit (Illumina) or ThruPLEX® DNA-seq Kit (Rubicon Genomics) for next-generation sequencing (NGS) on Illumina HiSeq 2500 platform.

Chromatin immunoprecipitation sequencing (ChIP-seq) [55]

For the ChIP-seq of phosphorylation of histone variant H2AX on S139 (γ-H2AX ChIP-seq) [118], formaldehyde was added to the culture medium to a final concentration of 1% for 10 min at room temperature. Glycine was added to a final concentration of 0.125 M for 5 min to stop crosslinking. Cells were harvested by scraping after PBS wash. Pelleted cells were lysed in lysis buffer (50 mM PIPES, pH 8; 85 mM KCl; 0.5% NP-40). The lysates were homogenized with a Dounce homogenizer and nuclei were harvested by centrifugation. Nuclei were then incubated in nuclear lysis buffer (50 mM Tris, pH 8.1; 10mM EDTA; 1% SDS) and sonicated at 70% amplitude for a duration of 3 min and 25 s with 15 s on and 45 s off (Qsonica Q700 sonicator) to obtain DNA fragments of about 500-1000 bp. Samples were diluted 10 times in dilution buffer (0.01 % SDS; 1.1% Triton X-100; 1.2 mM EDTA; 16.7 mM Tris, pH 8.1; 167 mM NaCl) and subjected to a 45 min preclearing with 140 µl of previously blocked protein-A and protein-G beads. Blocking was achieved by incubating the agarose beads with 500 µg of BSA and 200 µg of herring sperm DNA for 3 h at 4 °C. Precleared samples were incubated overnight at 4 °C with antibodies specific for γ- H2AX (10 µl) or without antibody as negative control. Immune complexes were then recovered by incubating the samples with 140 µl of blocked protein-A/protein-G beads for 2 h at 4 °C on a rotating wheel. Beads were washed once in dialysis buffer (2 mM EDTA; 50mM Tris, pH 8; 0.2% Sarkosyl) and four times in wash buffer (100 mM Tris, pH 8.8; 500 mM LiCl; 1% NP-40; 1% NaDoc). Elution from the beads was achieved by incubation in elution buffer (1% SDS; 100mM NaHCO3) for 15 min. Crosslink was reversed by adding 0.2% SDS and RNase A to the samples and incubating overnight at 70 °C. After a 2-h proteinase K treatment, DNA was precipitated by phenol/chloroform extraction and ethanol precipitation. The AsiSI–ER-U2OS cells treated with or without 4-hydroxytamoxifen (4-OHT) were included as positive control for the validation of experiments. The pulled down material and input DNA were then size-selected, and ligated to Illumina barcoded adaptors, using TruSeq ChIP Sample Preparation Kit (Illumina) or ThruPLEX® DNA-seq Kit (Rubicon Genomics) for next-generation sequencing (NGS) on Illumina HiSeq 2500 and HiSeq 4000 platforms. For phospho-RPA2-S33 ChIP, similar procedure was performed with minor modifications. Cells were resuspended in sonication buffer (50 mM HEPES, pH 8.0; 140 mM NaCl; 1 mM EDTA; 1% Triton X-100; 0.1% NaDoc; 0.5% SDS) and proceeded to sonication. Immunoprecipitation was performed using 30 µg chromatin and 4 µg anti-phospho-RPA2-S33 antibody. The pulldown material was eluted using IPure kit (Diagenode) and proceeded to NGS as described above.

Immobilized-Breaks Labeling, Enrichment on Streptavidin and next-generation Sequencing (i-BLESS)

Samples for i-BLESS analysis were prepared as described [119] with minor modifications. Approximately 10 million of HeLa cells were resuspended in PBS buffer and mixed with 1% low melting point agarose in PBS buffer at 40 °C. Cell suspension was mixed with liquid paraffin at 40 °C and vigorously shaken by hand for 1 min, until emulsion was formed. The emulsion was then poured into ice-cold PBS buffer and the mixture was stirred for several minutes. Agarose bead sus- pension was gently centrifuged (200 × *g*, 10 min), paraffin layer was removed and agarose bead pellet was washed 3 times with TE buffer. Beads were washed with ES buffer (1% Sarkosyl, 25 mM EDTA, pH 8.0), resuspended in ES with 50 µg ml^{-1} Proteinase K and incubated overnight at 50 °C. After incubation, the beads were washed with TE + 0.1 mM PMSF and twice with TE. Next, the beads were washed in 1 × Blunting Buffer (NEB), followed by DNA ends blunting using Quick Blunting kit (NEB) for 2 h and then washed twice with TE. The beads were subsequently washed with dA-Tailing Reaction Buffer (NEB) and DNA ends were A-tailed using NEBNext® dA-Tailing Module for 80 min. Next, the beads were washed with T4 ligation buffer and then resuspended in T4 ligation buffer with 100 nM P5 adapter and T4 ligase (NEB) and incubated overnight at 16 °C. After ligation, the beads were washed once with TE, and encapsulated DNA was initially.

RNA-seq

RNA-seq libraries were prepared using the Illumina TruSeq Stranded mRNA Library Prep Kit. Paired-end RNA-seq were performed with an Illumina NextSeq sequencing instrument (Helixio, France).

Comet assay

DNA breaks were monitored using the OxiSelect Comet Assay Kit (CELL BIOLABS, Inc.) according to the manufacturer's instructions. Slides were visualized using immunofluorescence microscopy (Zeiss ApoTome). The acquired comet images were analyzed by using MetaMorph Microscopy Automation and Image Analysis Software (Molecular Devices) and statistical analysis was performed with GraphPad Prism (GraphPad Software). A total of 200 cells were analyzed.

Available published datasets

Global nuclear run-on sequencing (GRO-seq) [120] for detecting the nascent transcripts and OK-seq [12] for detecting the replication fork direction and initiation/termination zones are obtained directly from accession GSE62046 and SRP065949 respectively.

1.3. Bioinformatics analysis

The quality of sequencing data was checked with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). ChIP-seq and DRIP-seq data were aligned to Human genome reference (hg19 assembly) with Bowtie2 [121] and RNA-seq using STAR [122]. Mapping quality was assessed with SAMtools [123] and in-house Python scripts. Peak-calling for DRIP-seq data was done using MACS2 [124] with a q-value of 0.05 and keeping up to five duplicates. Only the expressed genes, with the transcription RPKM>0, were selected to determine the impact of different gene positions on R-loop formation. Intersection of transcripts annotation (RefSeq, hg19) with R-loop signal was done using BEDTools [125]. The analyses of replication fork directionality and replication initiation zones used the published OK-seq data from HeLa cells. DeepTools2 [112] was used to compute and draw enrichment heat maps and profiles on positions of interest (peaks, TSS, TTS). Further analyses were done in R (http://www.R-project.org), with Bioconductor [126] packages and ggplot2 for graphic representation.

2. Results

2.1. TOP1 depletion increases R-loop levels

TOP1 is essential for cell growth and an acute depletion of this enzyme leads to a G_0/G_1 arrest [127]. To monitor the effect of TOP1 depletion on TRCs, we constructed a stable HeLa cell line expressing an inducible shRNA against TOP1 (shTOP1). Conditions of depletion were optimized to reduce TOP1 levels without arresting cell cycle progression. This is confirmed by the fact that the distribution of cells in G_1, S and G_2 phases of the cell cycle was not affected by the reduction of TOP1 levels (Fig. 3-1A, B). To monitor the impact of this depletion on replication forks, cells were labeled for 20 minutes with 5-iodo-2'-deoxyuridine (IdU) and for 20 minutes with 5-chloro-2'-deoxyuridine (CldU). DNA fibers incorporated with halogenated thymidine analogs were detected by immunofluorescence using specific antibodies by DNA combing technique [43]. We observed a 30 to 40% reduction of CldU tracks length in TOP1-depleted cells relative to control cells, which was largely suppressed by a transient overexpression of human RNase H1 (Fig. 3-1C) which highly suggests that the replication

slowdown observed in shTOP1 cells is caused by RNA:DNA hybrids. To confirm that TOP1-depleted cells have increased levels of R-loops, we used the S9.6 monoclonal antibody to quantify RNA:DNA hybrids in control and shTOP1 cells and observed an 70% increase in R-loop levels in shTOP1 cells. Importantly, this signal was highly sensitive to RNase H, confirming thereby that it corresponds to RNA:DNA hybrids.

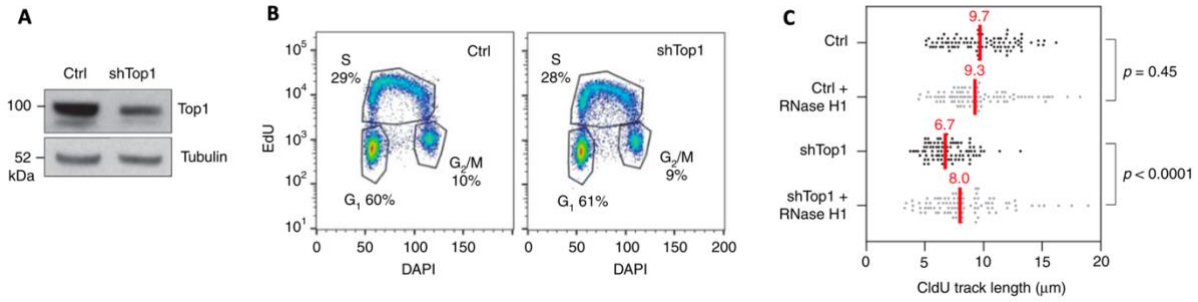

Figure 3-1. Depletion of TOP1 does not impact the cell cycle but slows down the fork progression. (A) Western blot analysis of TOP1 levels in HeLa cells expressing shRNA targeting TOP1 (shTOP1) under the control of a doxycycline-inducible promoter at 72 h post-induction (n = 9). (B) Cell-cycle distribution of control and shTOP1 cells determined by flow cytometry after labeling of S-phase cells with EdU. The fraction of cells in the different cell-cycle phases is indicated. (C) Doxycycline-treated control and shTOP1 HeLa cells were transfected for 48 h with a mock vector (EGFP-N1) or human RNase H1-EGFP (+RNase H1) and were sequentially labeled with IdU and CldU for 20 min. Replication fork progression was measured using DNA fiber spreading. The median length of CldU tracks is indicated in red. At least 150 fibers of each sample were measured (n = 3). P-values were calculated with the two-sided Mann–Whitney rank-sum test. Figures provided by our collaborators Pasero's lab.

2.2. R-loops form preferentially at TSS and TTS

To identify regions of the human genome that are prone to form R-loops in the absence of TOP1, RNA:DNA hybrids were immunoprecipitated with the S9.6 antibody and were analyzed by next generation sequencing (DRIP-seq) as described [81]. DRIP-seq profiles showed the enrichment of R-loops overlapped with 8726 and 10906 annotated genes (RefSeq annotations, hg19) in control and shTOP1 cells, respectively and most of R-loop positive genes (8015) were common to both cell types with high transcription level (Fig 3-2A). On average, they are mainly enriched at both TSS and TTS (Fig. 3-2B, C), which is consistent with the previous studies [80]. The validation of R-loop enrichment at TTS of genes was performed by DRIP-qPCR (Fig. 3-2D). Genes showed similar enrichment patterns of R-loops and expression in both control and shTOP1 cells. DRIP signals are further increased at TTS of converging genes in a manner that depended both on the distance between converging genes and on their level of expression (Fig. 3-2E, F). Together, these data indicate that the TSS and TTS of highly expressed genes

represent hotspots of R-loops and that shTOP1 cells show increased R-loop levels and slower fork progression.

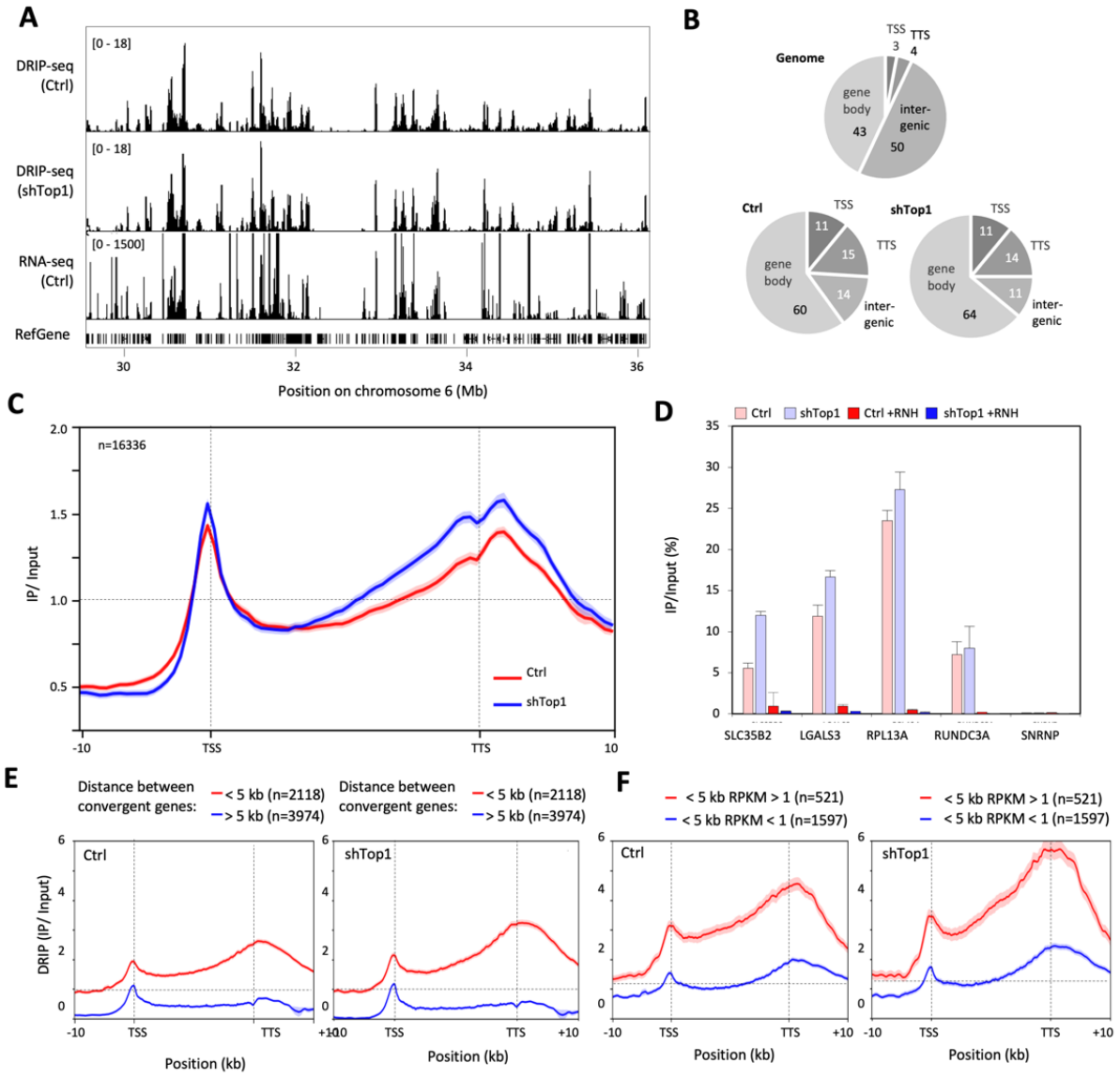

Figure 3-2. R-loop distribution in genome wide. (A) DRIP-seq expressed in RPKM (Read Per Kilobase per Million reads) in control and shTOP1 HeLa cells. A representative region on chromosome 6 is shown. RNA-seq data (RPKM) for HeLa cells and gene positions (hg19) are also indicated. (B) Distribution of R-loop peaks relative to gene annotations in control and shTOP1 HeLa cells. Peaks were obtained with MACS2 and were analyzed with CEAS (Cis-Regulatory Element Annotation System). The expected distribution in CEAS peaks were randomly positioned in the genome is shown for comparison. The percentage of DRIP-seq signals present in each annotation class is indicated. TSS: Transcription Start Site (5′-UTR and 3 kb upstream). TTS: Transcription Termination Site (3′-UTR and 3 kb downstream). (C) Metaplot of the distribution of S9.6 signals (IP/input) along 16,336 active human genes (RPKM > 0) and flanking regions (±10 kb) in control (red) and shTOP1 (blue) HeLa cells. Error bars correspond to SEM. (D) DRIP-qPCR analysis of the relative enrichment of RNA:DNA hybrids at the TTS of 4 genes and a negative control region (SNPRN) in control and shTOP1 HeLa cells after RNase H1 treatment (+RNH). Error bars correspond to three independent experiments. (E, F) Metaplots of DRIP signals at converging genes depending on the distance between their TTS in both conditions and at converging genes spaced by less than 5 kb relative to mRNA levels in control and shTOP1 HeLa cells. Shadows indicate standard error. Figure D provided from Pasero's lab.

2.3. Phospho-RPA accumulates at TTS of R-loop containing genes

To identify RNA:DNA hybrids that may interfere with fork progression, we next used the phosphorylation of RPA32 by ATR on S33 (p-RPA) as a surrogate for stalled replication and ATR activation. Regions enriched in p-RPA were mapped by ChIP-seq and were positioned relative to DRIP signals in untreated control cells (Fig. 3-3A). Same p-RPA patterns are also found in shTOP1 cells. The detection of individual DRIP and p-RPA peaks on DRIP-seq and ChIP-seq profiles (Fig. 3-3A; underlined in black) revealed that although most genes enriched in p-RPA contained R-loops, only a fraction of R loop containing genes were enriched in p-RPA in control and shTOP1 cells (Fig. 3-3B). 27% of all the R-loops detected in control cells and 7% of R-loops detected in shTOP1 cells are overlapped with p-RPA enriched regions indicating that most R-loops do not interfere with fork progression. However, up to 90% of detected p-RPA signals are also co-enriched with DRIP signals suggesting that R-loops are definitely involved in the replication fork stalling process.

To identify the R-loops that are potentially toxic for replication forks, we compared the distribution of DRIP signals and p-RPA at annotated genes. Unlike R-loops, p-RPA was mostly present at TTS and not at TSS in control and shTOP1 cells (Fig. 3-3C, D). More precisely, this is illustrated with the MED15 gene, which shows a peak of p-RPA downstream of TTS and no enrichment at TSS (Fig. 3-3A). These data suggest that forks preferentially pause at the TTS of highly expressed genes containing R-loops.

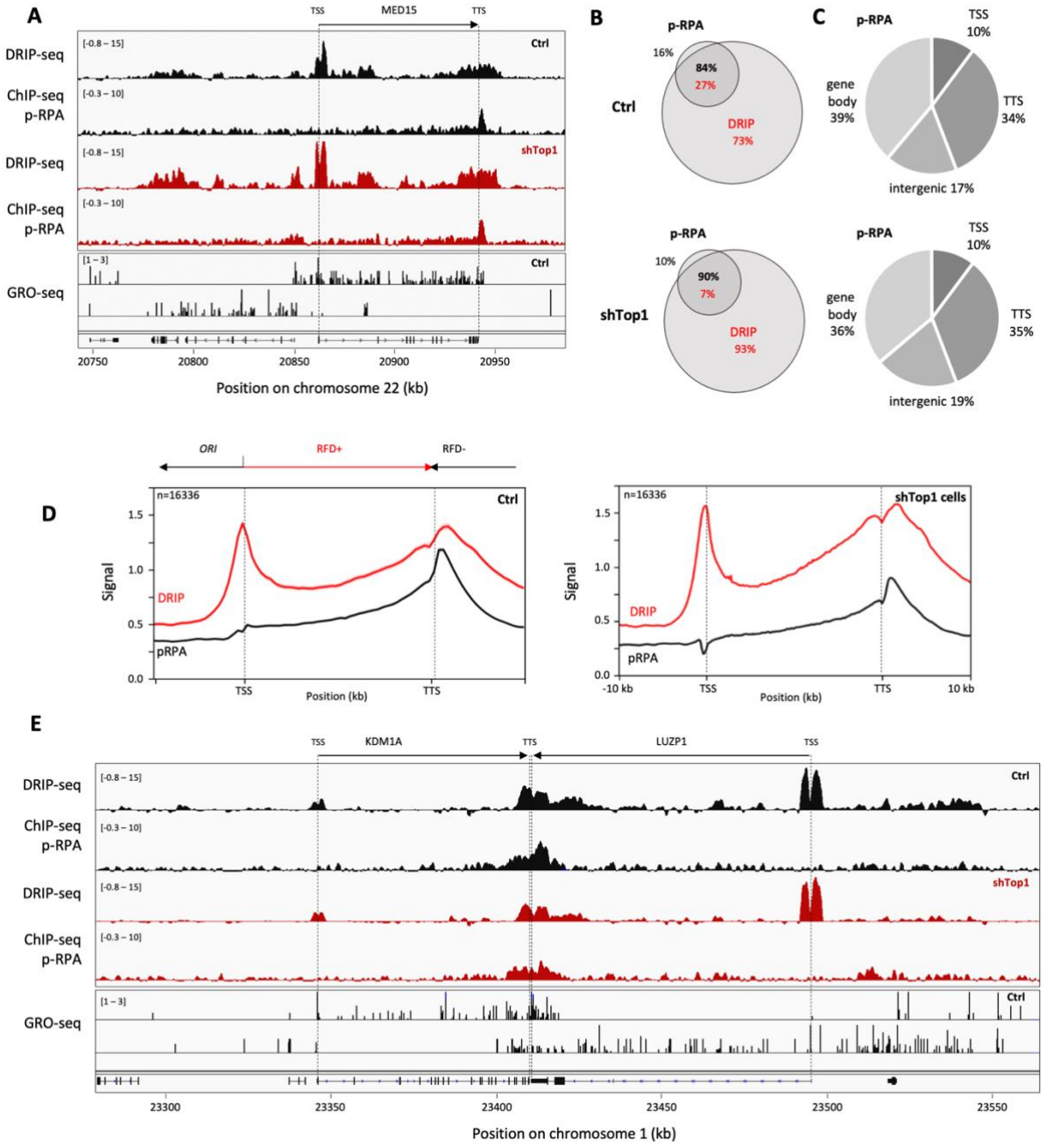

Figure 3-3. Phospho-RPA accumulates at TTS in control and shTOP1 cells. (A) Distribution of DNA hybrids (DRIP-seq), p-RPA32 S33 (ChIP-seq) and nascent transcription (GRO-seq) signals at a representative region on chromosome 22 in control HeLa cells and shTOP1 cells. The positions of TSS and TTS are indicated for the MED15 gene. The positions of DRIP and p-RPA peaks identified with MACS2 are also indicated. (B) Venn diagram of the percentage of genes overlapping with R-loop (red) and p-RPA peaks (black) peaks (MACS2) in control and shTOP1 cells. (C) The distribution of p-RPA peaks was analyzed with CEAS as in Fig 1e. The percentage of p-RPA peaks present in each region is indicated. (D) Metaplots of RNA:DNA hybrids (DRIP, red), p-RPA (black) and replication fork direction (RFD, blue) at 16336 active genes in control and shTOP1 cells. (E) Distribution of RNA:DNA hybrids (DRIP-seq), p-RPA32 S33 (ChIP-seq) and nascent transcription (GRO-seq) signals at two converging genes KDM1A and LUZP1 on chromosome 1 in control and shTOP1 HeLa cells. The positions of DRIP and p-RPA peaks identified with MACS2 are indicated.

2.4. Phospho-RPA accumulates at TTS in a head-on orientation

HO collisions between replication and transcription are generally considered more harmful than CD collisions [3,5]. Since highly expressed genes usually contain active replication origins in their promoter region and are therefore mostly replicated codirectionally with transcription [12,75], we reasoned that the asymmetric distribution of p-RPA at genes could reflect this bias in replication fork direction (RFD). To address this possibility, we analyzed the direction of fork movement at gene loci using Okazaki fragment sequencing data and the bioinformatics methods developed by our lab [12]. In Figure 3-4A, the MED15 gene contains a replication origin in its promoter region and is mostly replicated by forks progressing co-directionally with transcription. In contrast the TTS region of MED15 is preferentially replicated by an origin located downstream of the gene. Remarkably, p-RPA enrichment was detected at TTS, where replication and transcription converge, and not at TSS, which is replicated in a CD orientation. This p-RPA enrichment at TTS regions replicated in a HO orientation (RFD score < 0), but not at TSS replicated in a CD orientation (RFD score > 0) was also observed on a metaplot of 16336 active genes (Fig. 3-4B), indicating that it is a general feature of the human genome.

Since the TTS of converging genes are hotspots for RNA:DNA hybrids (Fig. 3-4B), we next exanimated whether it is also the case for p-RPA. As illustrated in Fig. 3-3E, p-RPA was enriched at the TTS of the converging genes KDM1A and LUZP1 in both control and shTOP1 cells. Phospho-RPA was also enriched at the TTS of 2118 converging genes separated by less than 5 kb, but not for 3974 TTS separated by more than 5 kb (Fig. 3-4C). The amount of p-RPA depended on the level of expression of converging genes (Fig. 3-4D), as it is the case for R-loops (Fig. 3-2F). Interestingly, p-RPA enrichment at TTS was also influenced by the presence of a nearby replication origin downstream of the TTS (Fig. 3-4E), similar to what was observed for the MED15 gene (Fig. 3-4A). The amount of p-RPA at TTS decreased as the distance between TTS and the replication origin increased (Fig. 3-4E), presumably because a short distance to the next downstream origin increases the risk of HO collisions at TTS (Fig. 3-4F; RFD score < 0). Altogether, these data indicate that the accumulation of p-RPA at TTS is determined by the direction of replication forks and gene transcription.

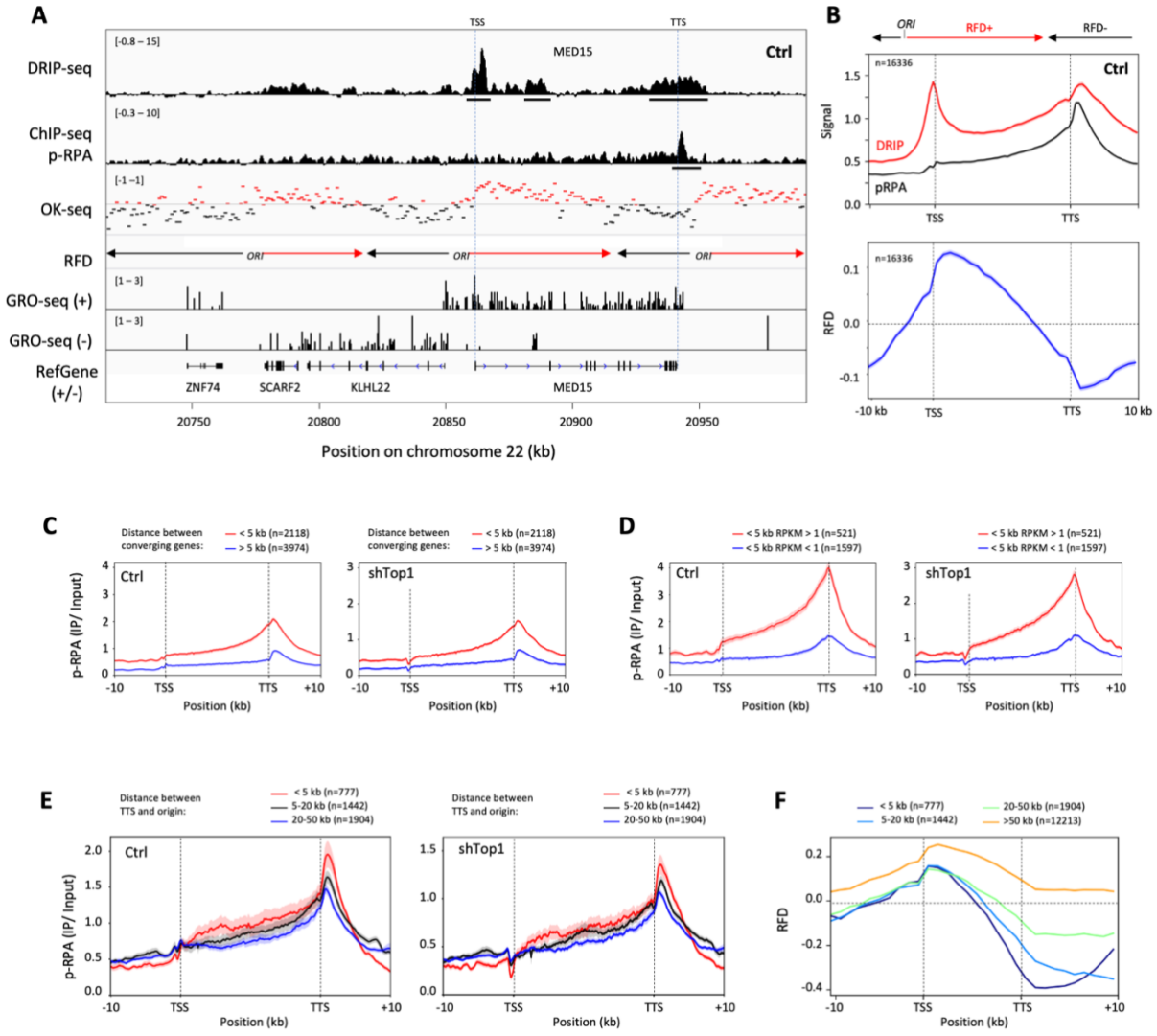

Figure 3-4. phospho-RPA accumulates at TTS in a head-on orientation. (A) Distribution of RNA:DNA hybrids (DRIP-seq), p-RPA32 S33 (ChIP-seq), Okazaki fragments (OK-seq) and nascent transcription (GRO-seq) signals at a representative region on chromosome 22 in control HeLa cells. Replication fork direction (RFD) is derived from OK-seq data. The positions of TSS and TTS are indicated for the MED15 gene. The positions of DRIP and p-RPA peaks identified with MACS2 are also indicated. (B) Metaplots of RNA:DNA hybrids (DRIP, red), p-RPA (black) and replication fork direction (RFD, blue) at 16336 active genes in HeLa cells. (C) Metaplots of the distribution of S9.6 signals at converging genes depending on the distance between their TTS in control and shTOP1 HeLa cells. (D) Metaplots of the distribution of S9.6 signal at converging genes spaced by less than 5 kb relative to mRNA levels in control and shTOP1 HeLa cells. Shadows indicate standard error. (E) Metaplots of p-RPA signals at genes in control and shTOP1 cells depending on the distance between TTS and the next downstream replication origins, shadows indicate standard error. (F) Impact on replication fork direction (RFD) of the distance between genes and downstream origins.

2.5. TOP1-depleted cells accumulate γ-H2AX and DSBs

To further characterize the impact of R-loops on replication stress and chromosome breaks, we next analyzed the presence of γ-H2AX in control and shTOP1 cells. Western blot analyses revealed a global increase of γ-H2AX levels in shTOP1 cells relative to control cells (Fig. 3-

5A). This is consistent with an increase of spontaneous DNA breaks in shTOP1 cells relative to control cells, as determined by comet assay (Fig. 3-5B) and to an increase of p-RPA32 S4/S8 foci in shTOP1 cells, which is indicative of DSBs (Fig. 3-5C). To determine whether γ-H2AX accumulates at the TTS of highly expressed genes containing R-loops in the absence of TOP1, the 35251 annotated genes were sorted according to their mRNA level (RPKM) and were organized in five quintiles (7050 genes) of decreasing gene expression level (Fig. 3-5D). The analyses of DRIP-seq and ChIP-seq signals at TTS revealed that although the distribution of R-loops and p-RPA was nearly identical in control and shTOP1 cells, γ-H2AX was only detected in shTOP1 cells. This signal decreased with gene expression level, which parallels the decrease of DRIP-seq and p-RPA ChIP-seq signals in the same quintiles.

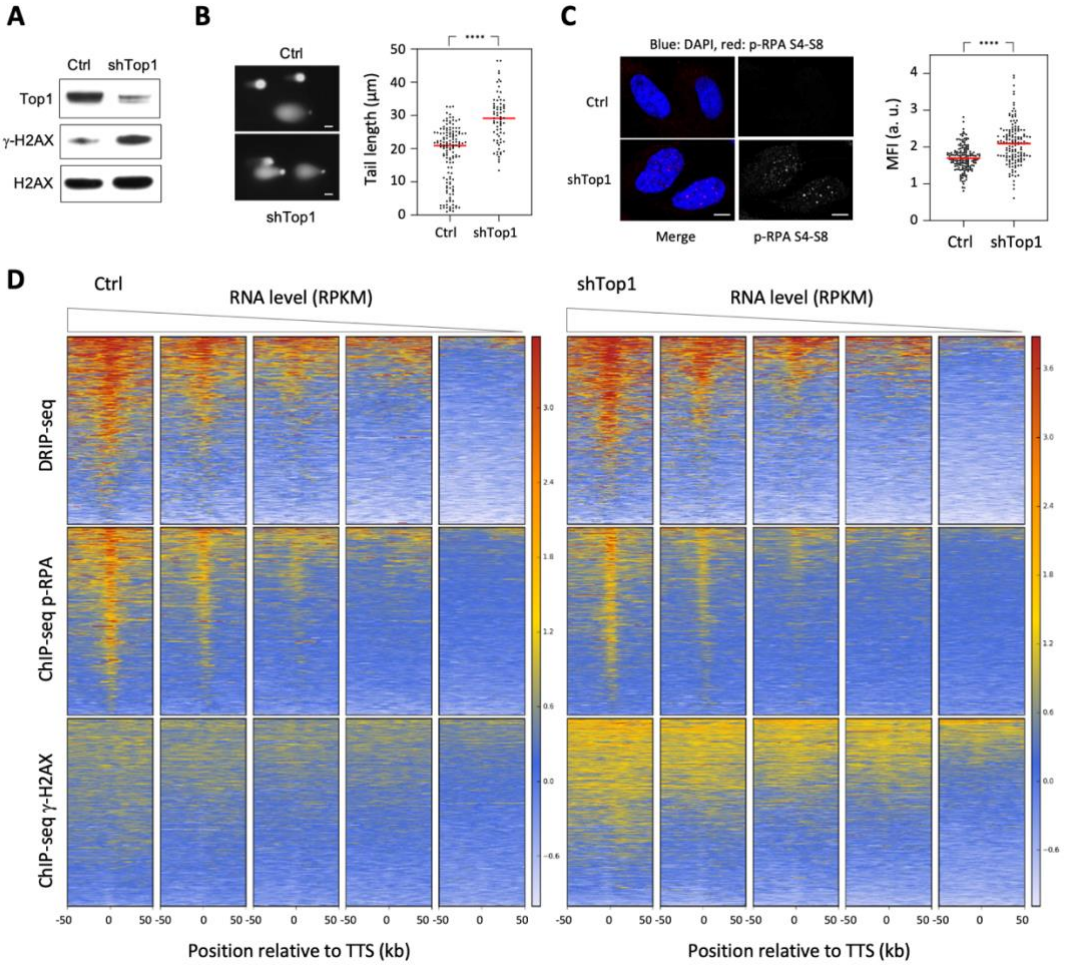

Figure 3-5. TOP1 prevents the accumulation of γ-H2AX at highly expressed genes. (A) Western blot analysis of g-H2AX levels in control and shTOP1 cells. (B) Analysis of DNA breaks in control and shTOP1 cells. Representative images and the distribution of comet tail lengths are shown. Bar is 10 µm. ****: p<0.0001, Mann-Whitney rank sum test. (C) Immunodetection of phospho-RPA32 S4/S8 in control and shTOP1 cells. Mean fluorescence intensity (MFI) of the p-RPA32 S4/S8 signals is shown. ****: p<0.0001, Mann-Whitney rank sum test. (D) Heat map of the intensity of RNA:DNA hybrids (DRIP), p-RPA and γ-H2AX at TTS in control and shTOP1 HeLa cells for five groups of genes with decreasing expression levels (RNA-seq). In each group, genes were sorted relative to the intensity of DRIP signal. Figure A, B and C are provided from Pasero's lab.

2.6. SRSF1-deficient cells do not phenocopy shTOP1 cells defects

The accumulation of DSBs and γ-H2AX in shTOP1 cells could either be due to R-loops or to topological stress. To discriminate between these possibilities, we depleted the splicing factor SRSF1 in HeLa cells to increase the formation of R-loops without affecting DNA supercoiling [128,129]. We observed an increased level of R-loops at highly expressed genes in shSRSF1 cells, showing the same distribution as in control and shTOP1 cells (Fig. 3-6A, B). Interestingly, shSRSF1 cells also showed an accumulation of p-RPA at TTS of highly expressed genes containing R-loops (Fig. 3-6C-G), but no increase in γ-H2AX levels (Fig. 3-6F-I), unlike shTOP1 cells (Fig. 3-5D). Altogether, these data suggest that DSBs form more frequently in shTOP1 cells than in shSRSF1 and control cells, which revealed that TOP1 prevents breaks by resolving topological constraints at TTS and increased R-loops at TTS is necessary but not sufficient for DSB induction.

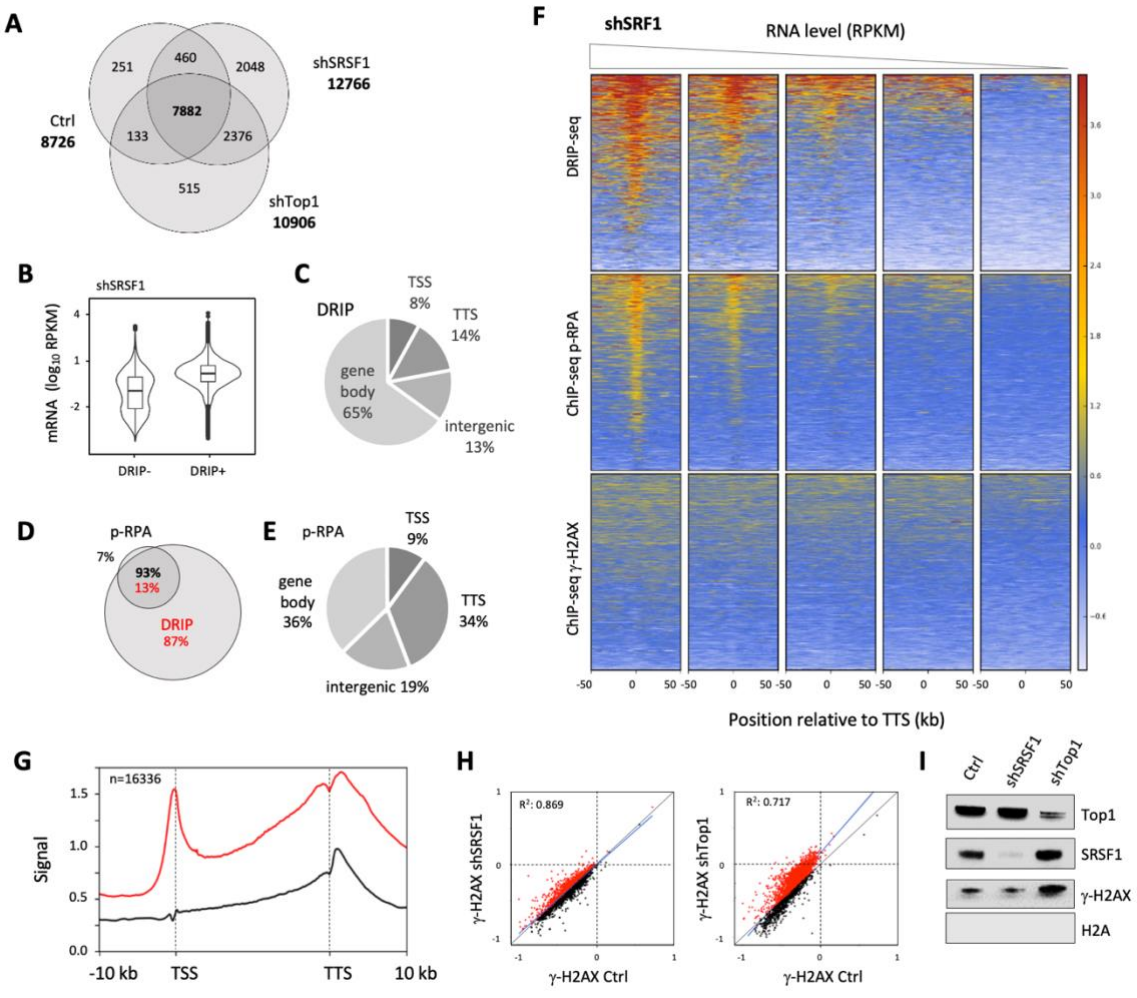

Figure 3-6. Depletion of SRSF1 increases R-loop and p-RPA at TTS, but not γ-H2AX. (A) Venn diagram of the number of genes enriched in R-loops in control, shSRSF1, and shTOP1 cells. R-loop-positive genes correspond to genes overlapping with R-loop peaks identified with MACS2. (B) mRNA level (RPKM) of genes overlapping (R-loop+) or not (R-loop−) with S9.6 peaks in shSRSF1 cells. Box: 25th and 75th percentiles; central line: median. (C) Distribution of R-loop peaks in shSRSF1 cells relative to gene annotations. Peaks were obtained with MACS2 and were analyzed with CEAS (Cis-Regulatory Element Annotation System). (D) Venn diagram of the percentage of genes overlapping with R-loop (red) and p-RPA peaks (black) peaks (MACS2) in shSRSF1 cells. (E) Distribution of p-RPA (S33) peaks in shSRSF1 cells relative to gene annotations. (F) Heatmap of the intensity of RNA:DNA hybrids (DRIP), p-RPA, and γ-H2AX at TTS in shSRSF1 cells for five groups of genes with decreasing mRNA levels (RPKM). Within each group, genes were sorted relative to the intensity of DRIP signal. (G) Metaplot of RNA:DNA hybrids (DRIP, red) and p-RPA (black) at 16,336 active genes in shSRSF1 cells. Error bars correspond to SEM. (H) Scatter plot of the intensity of γ-H2AX signal at all active genes in control, shSRSF1, and shTOP1 cells. (I) Western blot analysis of γ-H2AX levels on chromatin in control, shSRSF1, and shTOP1 cells. H2AX used as a loading control (n = 3). Figure I provided from Pasero's lab.

2.7. DSBs form at TTS containing R-loops in shTOP1 cells

Since the resolution of γ-H2AX ChIP-seq profiles is not sufficient to position chromosome breaks, we next used a next-generation sequencing-based assay called i-BLESS to map DSBs at nucleotide resolution. To determine whether shTOP1 cells accumulate DSBs at TTS, we measured the intensity of i-BLESS signal for a 2 kb window centered on the TTS of all human genes and sorted them according to the intensity of this signal (Fig. 3-7A). Interestingly, the TTS of the top 25% genes also showed an increased level of DRIP and p-RPA (Fig. 3-7B). A similar result was obtained when we used a hierarchical clustering approach to identify genes with increased i-BLESS signal indicating increased frequency of DSBs (DSB+, n=9533) at their TTS in shTOP1 cells (Fig. 3-7C, 3-8C). Again, DSB+ genes also showed increased levels of R-loops, p-RPA and γ-H2AX relative to DSB- genes (Fig. 3-7D). When we analyzed DSBs at TSS, we also found increased i-BLESS signal at a subset of genes in shTOP1 cells (Fig. 3-8A), which is consistent with the presence of transcription-dependent DSBs in promoter regions. However, these breaks were not associated with increased p-RPA levels (Fig. 3-8B), unlike at TTS (Fig.3-7A).

Finally, we analyzed the incidence of gene orientation on DSB formation. While the percentage of genes in converging (HO) or codirectional (CD) orientations was not significantly different between DSB+ and DSB- genes (44 vs 45% for HO), DSB+ genes showed increased DRIP and p-RPA signals at closely arranged HO genes (<5 kb between TTS) compared to DSB- genes (Fig. 3-8D). This difference was less marked for CD genes (Fig. 3-8E). Altogether, these data indicate that the increased γ-H2AX signal observed in shTOP1 cells results from DSBs occurring at the TTS of a large number of genes enriched in R-loops and p-RPA.

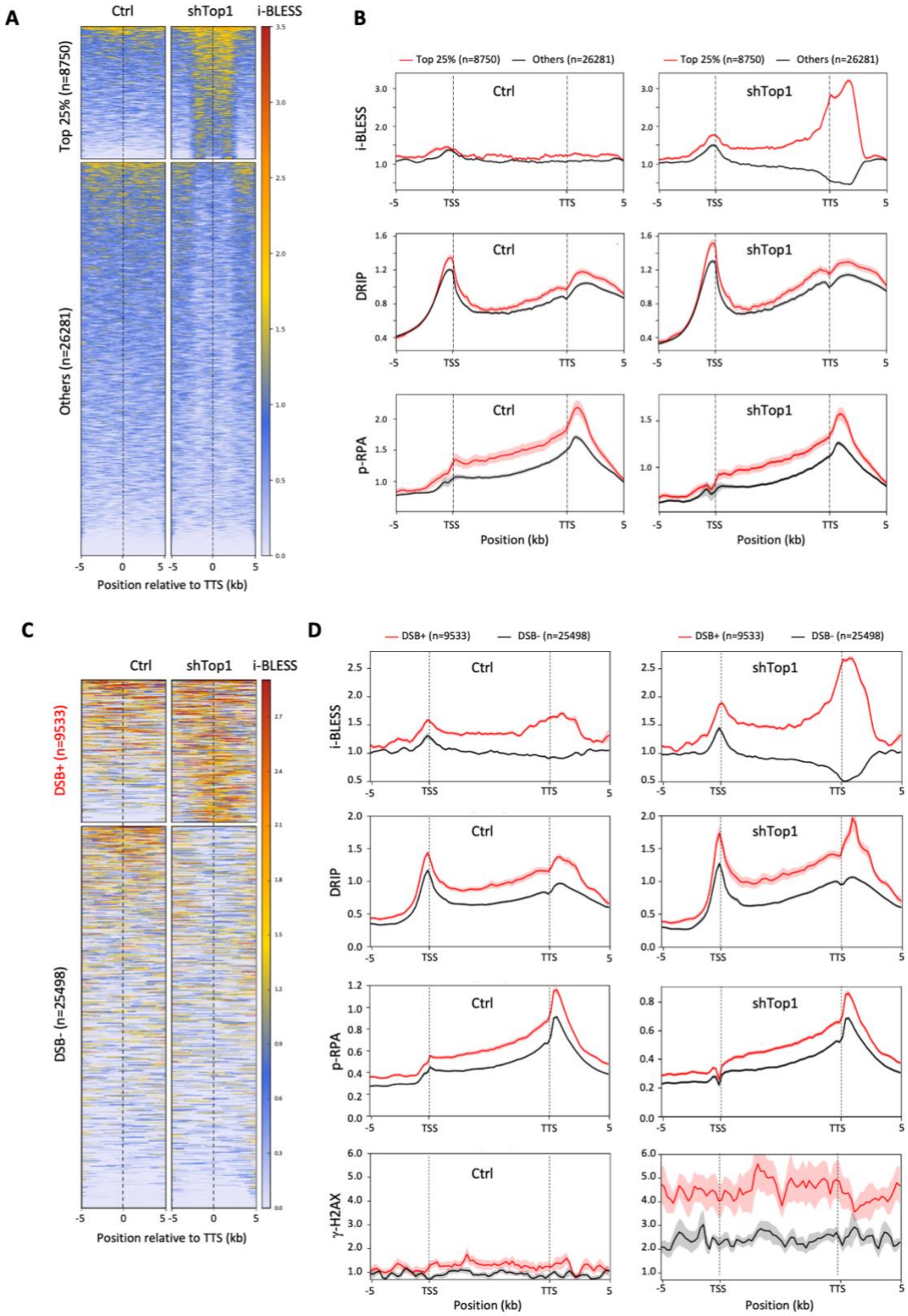

Figure 3-7. TOP1 prevents the accumulation of γ-H2AX and DSBs at TTS. (A) Heatmap of the intensity of i-BLESS signal at TTS in control and shTOP1 cells for two groups of genes determined according to the intensity of i-BLESS signal at the TTS (+/- 2 kb) in shTOP1 cells. (B) Metaplots of i-BLESS, RNA:DNA hybrids and p-RPA32 S33 signal for the Top25% (red) and others (black) genes in control and shTOP1 HeLa cells, shadows indicate standard error. (C) Heatmap of the intensity of i-BLESS signal at TTS in control and shTOP1 cells for two groups of genes (DSB+ and DSB-) determined by hierarchical clustering analysis of i-BLESS signal at the TTS in shTOP1 cells. (D) Metaplots of i-BLESS, RNA:DNA hybrids, p-RPA32 S33 and γ-H2AX signal for DSB+ (red) and DSB- (black) genes in control and shTOP1 cells, shadows indicate standard error.

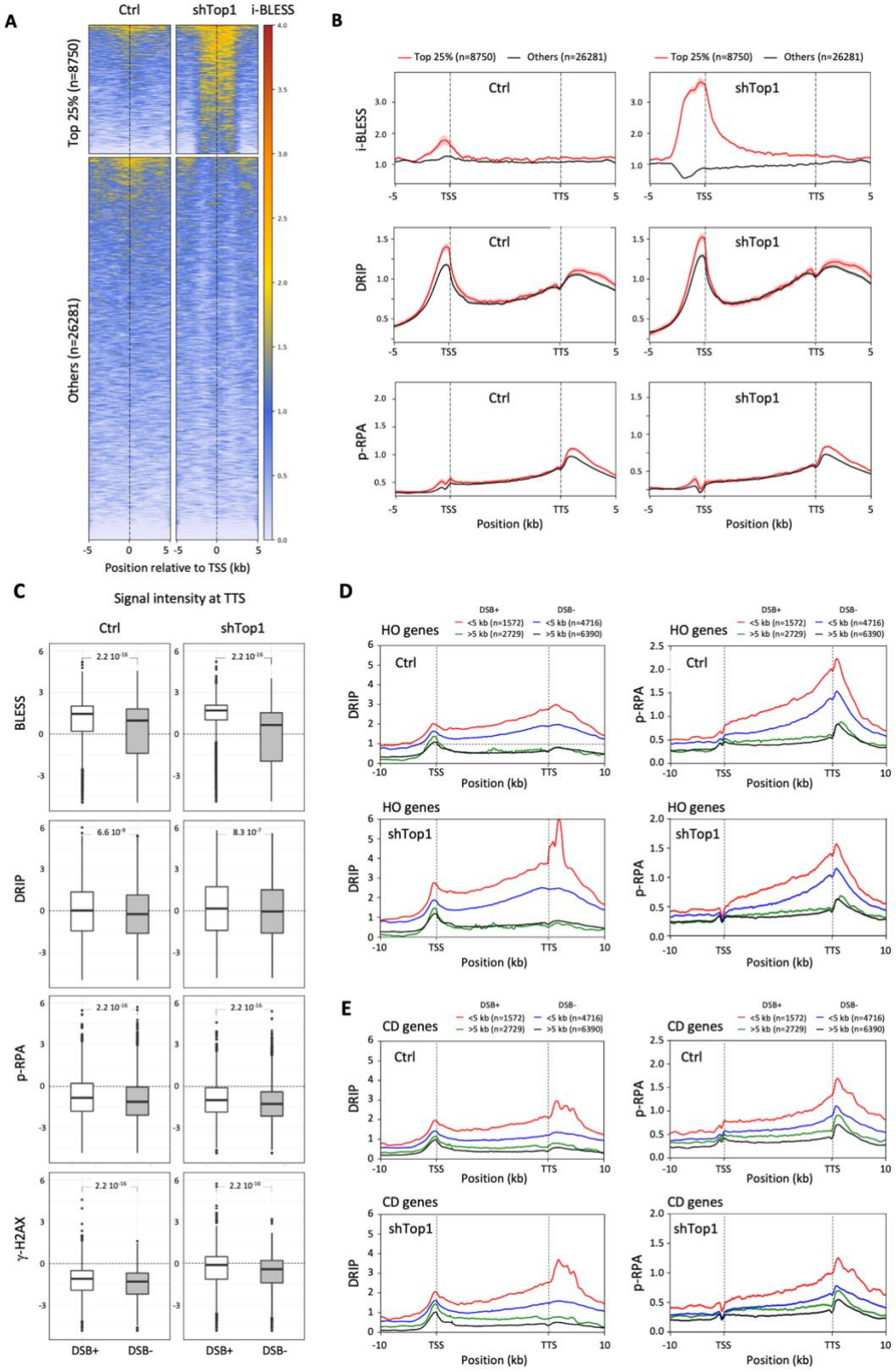

Figure 3-8. TOP1 prevents DNA breaks at TTS. (A) Heatmap of the intensity of i-BLESS signal at TSS in control and shTOP1 cells for two groups of genes determined according to the intensity of i-BLESS signal at the TSS (+/- 2 kb) in shTOP1 cells. (B) Metaplots of i-BLESS, RNA:DNA hybrids and p-RPA32 S33 signal for the Top25% (red) and others (black) genes in control and shTOP1 HeLa cells, shadows indicate standard error. (C) Distribution of i-BLESS, DRIP, p-RPA, γ-H2AX signal intensities at TTS of DSB+ (white) and DSB- genes (grey) in control and shTOP1 cells. (D, E) Intensity of DRIP and p-RPA signals at converging (HO) and codirectional (CD) genes from DSB+ and DSB- genes and separated by more or less than 5 kb in control and shTOP1 cells.

3. Discussion

It is now well established that R-loops have both positive and negative impacts on genome activity, but the difference between physiological and pathological R-loops has remained unclear. Here, we have compared the distribution of R-loops, replication stress markers (p-RPA and γ-H2AX) and DSBs in HeLa cells to identify R-loops that are detrimental to DNA replication and activate ATR. Using DRIP-seq, we have identified hotspots of R-loop formation at the promoters and terminators of highly expressed genes, as described earlier [80]. Depletion of TOP1 further increased R-loop levels at TTS and especially at converging genes, presumably because of the accumulation of topological stress [130,131]. Interestingly, we found that only 27% of R-loop containing genes colocalized with phospho-RPA32 (S33), a mark of ATR activation used here as a proxy for stalled replication forks. Yet, 84 to 90% of these p-RPA peaks were associated with R-loops. These values are derived from the conservative analysis of a weak ChIP signal in a population of unchallenged and asynchronously growing cells, so it could be that the actual number of p-RPA peaks is higher. Yet, these data indicate that p-RPA does not accumulate at all R-loops and suggest that only a fraction of R-loop containing genes are responsible for most of the fork pausing events in unchallenged growth conditions. Incidentally, these data indicate that the vast majority of the co-transcriptional R-loops present in the human genome do not interfere with DNA replication or at least do not induce a detectable activation of ATR.

One of the most striking differences between the distribution of DRIP and p-RPA signals is that R-loops were detected at both TSS and TTS of highly expressed genes whereas p-RPA was mostly enriched at TTS. Since promoter regions of highly expressed genes usually contain active replication origins [12,75], this asymmetry in p-RPA distribution may reflect an influence of fork polarity on transcription-replication conflicts [73,98]. A meta-analysis of replication fork direction through 16336 active genes (RPKM>1) confirmed that TSS and gene bodies are preferentially replicated codirectionally (RFD+), whereas TTS are mostly replicated by head-on forks (RFD-). Remarkably, p-RPA was enriched at RFD- regions, supporting the view that RPA is phosphorylated by ATR upon fork pausing at TTS enriched in R-loops. Our data are consistent with a recent study showing that R-loops interfere with fork progression in an orientation-dependent manner on a human episomal system [3] and extend this observation to the genome-wide level. Interestingly, p-RPA enrichment was further increased at the TTS of converging genes, proportionally to the levels of gene expression and to the proximity of the nearest HO-orientation gene neighbor. In addition, p-RPA levels at TTS were increased by the

proximity of a replication origin. Altogether, these data suggest that transcription terminators represent hotspots of R-loops and replication fork arrest in the human genome, acting in a context- and orientation-dependent manner.

TOP1 depletion in HeLa cells increased levels of γ-H2AX, phospho-RPA32 (S4/S8) and DNA breaks relative to control cells. To determine whether chromosome breaks occur at TTS, we have analyzed the distribution of DSBs at the nucleotide resolution using i-BLESS [119]. DSBs were detected downstream of the TTS of a large number of genes that were also enriched in R-loops and p-RPA, especially in regions of the genome where transcription converges. Since it has been recently reported that replication forks blocked by R-loops can be restarted by fork cleavage in a MUS81-dependent manner [132], an attractive possibility could be that DSBs detected at TTS are generated by this structure-specific endonuclease. Interestingly, DSBs were also detected upstream of TSS, which may correspond to the replication independent DSBs reported at promoter regions in other studies [133]. Recent reports indicate that these DSBs may depend on Topoisomerase II β activity and on the proximity of CTCF sites at loop anchors [134]. These breaks could be distinct from the replication-dependent DSBs occurring at TTS, which could be more related to the estrogen-induced DSBs occurring during S phase at R-loop-containing genes in breast cancer cells [88].

An important question that remains to be addressed is the mechanism by which R-loops interfere with DNA replication in human cells. It is generally proposed that RNA:DNA hybrids are intrinsically difficult to replicate and impede fork progression in an orientation-dependent manner. However, our DNA fiber analyses revealed that all replication forks were equally slowed down by 30 to 40% in shTOP1 cells, which argues against a direct effect of R-loops. Indeed, highly expressed genes cover only a small fraction of the human genome and R-loops should therefore affect only a subset of forks in shTOP1 cells. This should lead to a bimodal distribution of CldU track lengths and not a global reduction of fork speed. We rather favor a model in which replication fork pausing at TTS prevents HO collisions with transcription (Fig. 3-9). TOP1-deficient cells could experience difficulties to stabilize these paused forks, which would increase the risk of fork collapse and DSB formation. DSBs would in turn induce a chronic activation of S-phase checkpoints and a slowdown of replication forks. This view is supported by the fact that cells depleted for the splicing factor SRSF1, which have increased R-loop levels but no DNA relaxation problems, have faster replication forks and less γ-H2AX than shTOP1 cells. This model is also consistent with reports showing that ATR downregulates elongation at undamaged forks in yeast and human cells [135–137]. It is also consistent with data in

budding yeast showing that fork arrest does not directly depend on R-loops and is mechanistically separable from the induction of DNA damage [138]. Yet, the overexpression of RNase H1 partially rescued the slow fork phenotype of shTOP1 cells, suggesting that RNA:DNA hybrids negatively impact DNA replication in these cells. To explain this apparent discrepancy, we propose that RNA:DNA hybrids form at stalled forks as a consequence of fork arrest and could impede fork restart. This would be reminiscent of the formation of RNA:DNA hybrids at DSBs, which interfere with their HR-mediated repair [139].

In conclusion, our results suggest that polar fork arrest at TTS is an active process that prevents collisions between RNA and DNA polymerases, as previously reported in budding yeast [140,141]. Transient fork pausing could help cells displace RNA polymerases ahead of the replisome, through a process involving Mec1 and INO80 [142,143]. Since transcription is a discontinuous process [144], forks may also pause during transcription bursts and restart after passage of RNA polymerase convoys. In this model, TTS could act as traffic lights, arresting forks until road blocks have been removed. Alterations of DNA relaxation or pre-mRNA cleavage could perturb this coordination, leading to increased DNA breaks and to the chronic activation of ATR [128,145], which would reduce in turn the speed of replication forks. Our data are consistent with recent models in which initiation of DNA replication upstream of highly expressed genes would facilitate the coordination between replication and transcription [12]. This is reminiscent of the codirectional organization of genes in *B. subtilis* and other bacteria to avoid head-on conflicts with replication [98,146]. This organization does not exist in budding yeast, in which persistent R-loops were recently shown to cause genomic instability independently of their orientation [147]. In metazoan, this organization would accommodate extensive changes in gene expression profiles occurring during cell differentiation. In other words, the functional coupling between strong origins and promoters would represent a simple and flexible mean to limit transcription-replication conflicts in differentiating cells. Interestingly, it has been recently reported that the deregulation of oncogenic pathways activates intragenic replication origins that induce HO conflicts and chromosome breaks [72]. It is therefore tempting to speculate that the loss of a functional organization restraining replication-transcription conflicts to TTS leads to genomic instability in precancerous lesions.

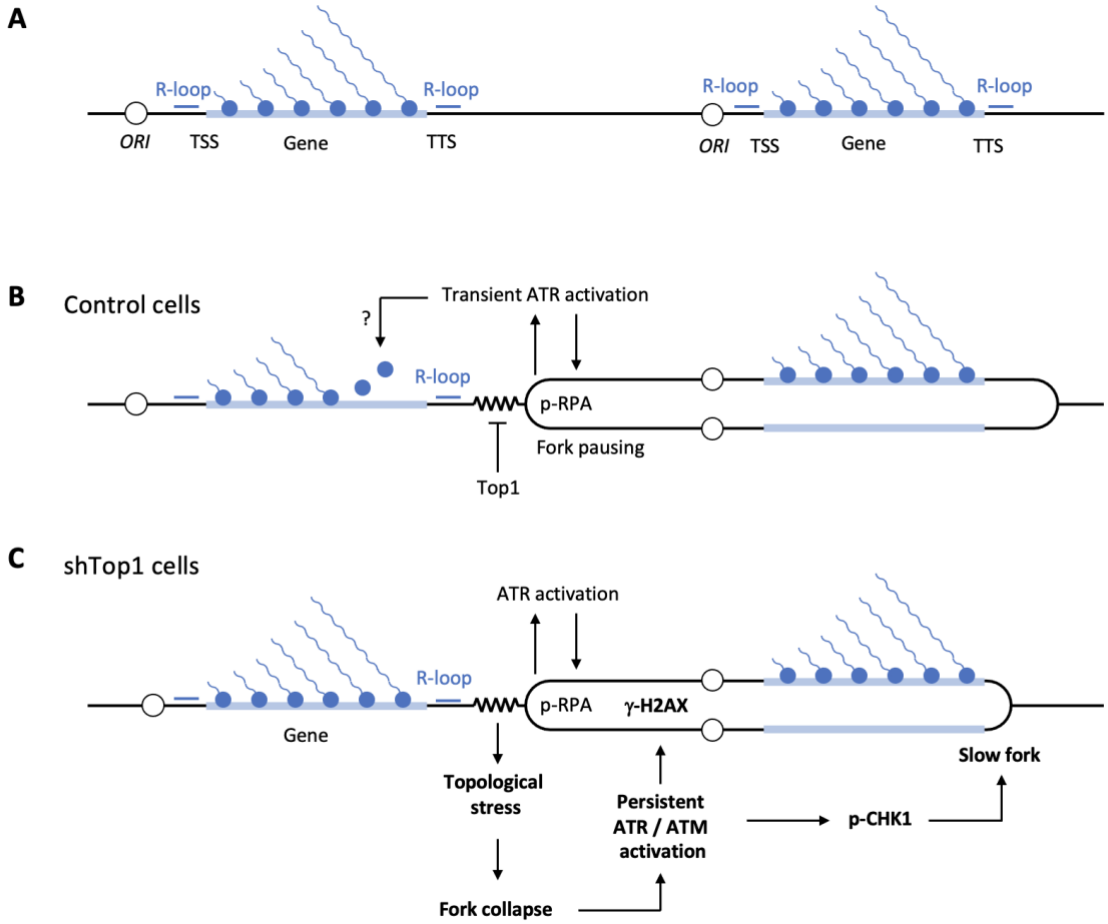

Figure 3-9. A two-step model for the regulation of TRCs in the human genome. (A) Highly expressed genes form co-transcriptional R-loops at TSS, TTS and to a lesser extent in gene bodies. Replication origins are frequently located upstream of TSS. (B) Initiation from upstream origins ensures that R-loops at TSS and gene bodies are preferentially replicated co directionally, which would limit HO conflicts. Forks progressing in the opposite direction pause when they encounter the TTS of highly expressed genes, presumably because of the accumulation of positive supercoiling. Transient fork pausing activates ATR and leads to the phosphorylation of RPA32 on S33. ATR may also promote the displacement of RNA polymerases ahead of the paused fork. (C) In the absence of TOP1, the accumulation of torsional stress may lead to fork collapse and to the sustained activation of ATR/ATM. This would in turn slow down fork progression throughout the genome.

Chapter 4 Perspectives

1. R-loop-induced transcription-replication conflicts (TRC) in breast cancer

Since TRC under replicative stress has been revealed to be a detrimental source in the homeostasis of genomic stability, strong evidences in recent decades also suggest that it could have a potential role in the tumor progression and cancer development [2]. Breast cancer is the most frequent malignancy occurring in females, accounting for ~25% of the total diagnosed cancer cases and 14% of the cancer deaths [148].

The estrogen receptor (ER) positive breast cancer accounted for around 75 % of all breast cancers, and increasing evidences from epidemiological, *in vivo* and *in vitro* studies suggest that the hormone estrogen (E2, 17b-estradiol) plays a causal role during carcinogenesis [149]. Specifically, higher E2 serum concentrations and longer lifetime E2 exposure are both positively correlated with an increased incidence of breast cancer. In addition, E2-mediated transcription induces DNA double-strand breaks (DSBs) specifically in breast epithelial cells that express the estrogen receptor in a cell cycle-dependent manner [150]. However, despite strong links between estrogen and genomic instability, the molecular mechanism by which E2 causes this instability in breast cancer is unclear.

Recent evidences have shed new light on the E2-dependent carcinogenic mechanism, showing that R-loops resulted from the E2 transcriptional response correlate with increased DNA damage in E2 dependent breast cancer [88]. Stork and colleagues have shown that, in MCF7 ER-positive breast cancer cells, the E2-induced transcriptional changes promote R-loop formation and DSBs. Importantly, it has been shown that such R-loop induced genome instability observed in MCF7 cells depends on DNA replication, as DNA damage is only detected in replicating cells, but not outside of the S phase (Fig.4-1A). Moreover, these E2-induced DNA damages were colocalized with R-loop formation (Fig. 4-1B, C) and could be reduced by R-loops degradation through RNase H overexpression [88]. Thus, one hypothesis to explain E2-induced genome instability is that the uncontrolled proliferation, due to too high/long E2 exposure, causes transcription-replication conflicts (TRC) that lead to replication stress and DNA damage.

To further optimize the identification of TRC loci, it would be interesting to compare replication fork progression (OK-seq) with transcription data derived from S-phase ER+ cells in presence and absence of E2. It will allow us to be able to identify more accurately the TRC distribution genome wide. Based on the previous research in HeLa cells, later combining with the R-loop (DRIP-seq), fork stalling (p-RPA) and DNA damage data (γ-H2AX or i-BLESS) in MCF7 cells with or without E2 treatment, it can finally determine the enriched loci and hotspots of gene where E2-induced TRC and R-loops in breast cancer that cause genome instability, which might favor the cancer prevention/therapy or drug development.

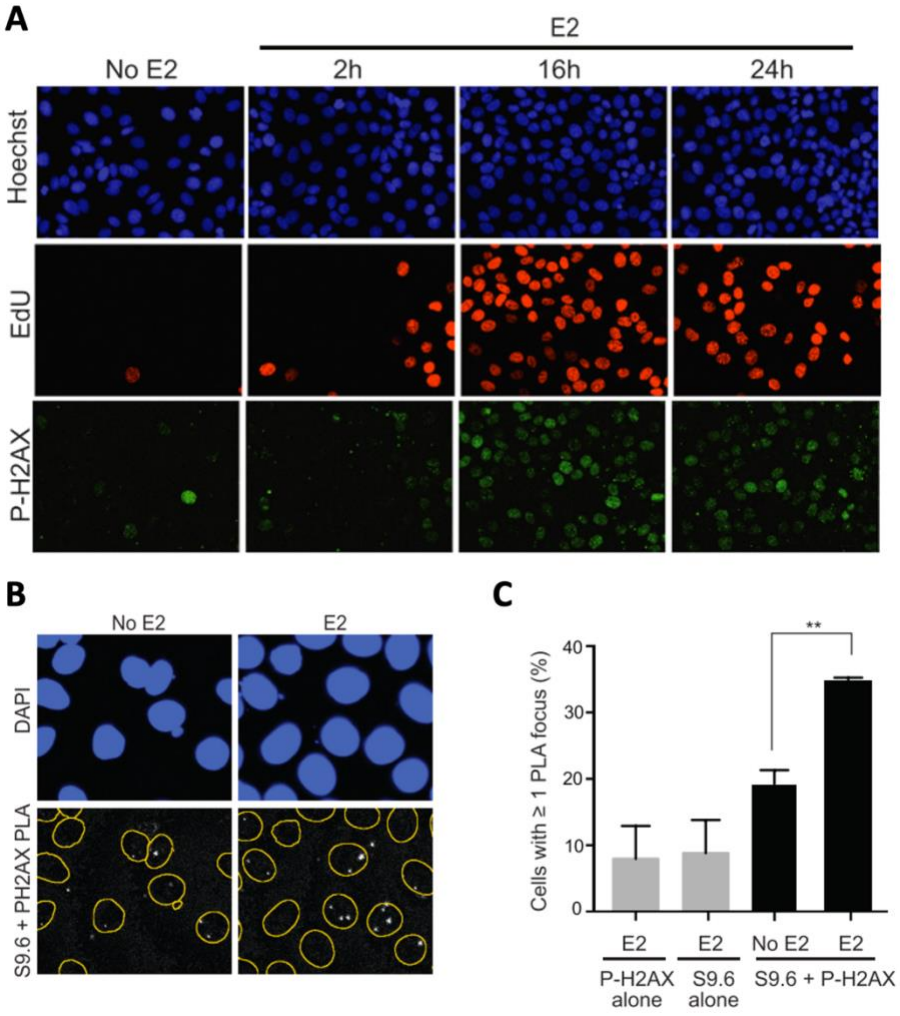

Figure 4-1. Estrogen induces R-loop formation and DNA damage in a replication dependent manner. (A) Immunostaining for EdU and P-H2AX in cells treated with 0 or 100 nM E2 for the indicated time (B) Proximity ligation assay (PLA) between S9.6 antibody (against R-loop) and P-H2AX antibody in cells treated with 100 nM E2 or not for 24 h. (C) Quantification of the percentage of cells with ≥ 1 PLA focus per nucleus. Single-antibody controls from cells treated with 100 nM E2 for 24 hr are shown. Error bars represent the SEM from 4 biological replicates. **p<0.01 (Student's t-test). Figrues adapted from Stork et al. 2016.

To further investigate the mechanism of co-transcriptional E2-induced R-loop involved in the TRC regulation and genomic stability maintenance in breast cancer, evidence shown that N^6-

methyladenosine (m6A) RNA modification and the main relative methyltransferase-like 3 (METTL3) could promote the breast cancer progression mediated also via R-loops [90,151,152]. Recently, as a new regulatory mechanism controlling gene transcription in eukaryotic cells, m6A, found in both mRNA and non-coding RNAs, plays a responsible role in almost all the bioprocesses including cancer pathogenesis [108]. Coincidentally, latest research revealed, m6A-related METTL3 may mediate R-loop formation connected to the m6A readers, e.g., YTHDC1, YTHDF2, around TTS to well regulate the transcription termination process [153] and generally promote homologous recombination repair of DSBs [90]. Besides, METTL3 is also found in breast cancer cells and tissue to potentially regulate the proliferation and apoptosis of cancer cells [107].

Combining the genome wide collected data on R-loops distribution, fork stalling and DNA damage together with TRC locations would absolutely shed light on the further study on the mechanism underline E2-induced genomic instability in breast cancer and provide new insight about possible new chemotherapy.

2. Study of R-loops in other diseases and cancers

As a genetic threat driving DNA damage and genome instability, the abnormal enrichment of R-loops has been well documented in many human diseases, cancers and syndromes. Collected data revealed that deregulating or mutating proteins involved in R-loop resolution can lead to severe diseases often associated with human neurological disorders [2,78]. Mutations in R-loop helicase SETX (also involved in DNA repair) is a major cause of amyotrophic lateral sclerosis type 4 (ALS4) and ataxia oculomotor apraxia type 2 (AOA2) (Fig. 4-2A) , two diseases characterized by a progressive degeneration of neurons in the brain and the spinal cord as well as muscle weakness [154]. Based on yeast data [155], it seems that SETX/Sen1 mutation can interfere with the regulation of transcription potentially leading to TRC events.

Other proteins involved in the R-loops resolution such as RNase H2, TREX1 (a ssDNA 3'-5' exonuclease), ADAR1 (a dsRNA-editing enzyme), and SAMHD1(dNTP triphosphatase) can be responsible, if mutated, for another neurological disease: the Aicardi–Goutières syndrome (AGS). This syndrome is characterized by an inflammatory disorder that mainly affects the bran and the skin (Fig. 4-2B) [156,157]. At the molecular level, depletion of RNase H2 or any of the other proteins above significantly induced a group of inflammation and immune-related mRNAs (IFNGR1, OAS1, TNF, STING, etc.,) with persistent R-loops accumulation and downregulation of transcription [158]. Other two diseases with the presence of an R-loop are

the Friedreich ataxia (FRDA) and the Fragile X syndrome (FXS). These are among 40 diseases linked to Triplet repeat expansions (TREs) [159]. In the specific case the expanding motif in FRDA is GAA_n while for the FXS is CGG_n. The expansion occurs in intergenic regions and depending on the number of repeated triplets the risk of forming a deleterious R-loop that can lead to DNA methylation-mediated silencing and disrupt the normal transcription regulation increases (Fig. 4-2C) [159]. Moreover, excluding Breast cancer, et least other four cancer types are linked to R-loop formation. In the eosinophilic leukemia, an oncogenic translocation inactivates the cleavage and polyadenylation factor FIP1L1, which promotes R-loop formation, leading to DNA damage and genomic instability (Fig.4-2D) [160]. In the Ewing's sarcoma a dysfunctional BRCA1 causes damage-induced transcription with accumulation of R-loops [161]. Burkitt's lymphoma and multiple myeloma are characterized by the fusion of the oncogene c-MYC with the immunoglobulin (Ig) locus [162–164]. Collected data suggesting that besides of the beneficial function of R-loops at facilitating the Ig class switch in normal condition [79], the abnormal chromosomal translocation might cause the rather detrimental R-loop formation in these cancer types. To conclude, understanding the link between R-loop, replication stress and genome instability in all these diseases and how to benefit from that to contribute to therapeutic methods development, need to be further investigated in future studies.

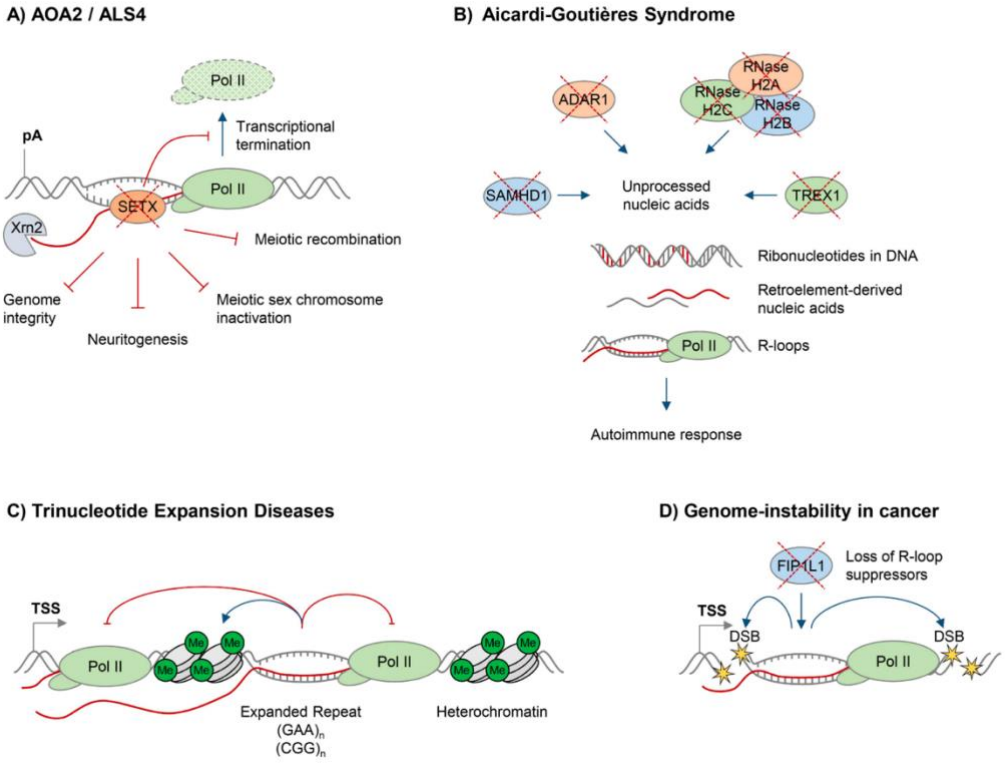

Figure 4-2. R-loops and human diseases. Each category shows a potential deregulation schema with loss of wild type protein function (by red crosses) or mutagenesis involved in R-loop formation in diverse diseases and cancers. Figure adapted from Groh et al. 2014.

3. Study of mutation landscape associated with TRC

As aforementioned, the replication program is routinely exposed to endogenous and exogenous stresses [2]. Evidences collected in recent years confirm that oncogene-induced replication stress is a major driver of mutagenesis and tumor progression [72,165].

To date, thousands of cancer mutational catalogs have been obtained covering almost all the existing cancer types [166]. This huge amount of data allows us to extract specific mutational patterns (also known as mutation signatures) resulted from processes of diverse cancers, e.g., base substitutions, small insertion and deletions (indels), genome rearrangement and chromosome copy number variation) [167]. These signatures are generally divided into transcriptional- and replicative-dependent asymmetric clusters. Transcriptional-related features are mainly associated with mutations originating from UV lights and smoking, whereas the replication-associated ones are linked to mutations on the genes encoding polymerase POLE, and APOBEC [100]. Previously, we have shown that R-loops are preferentially enriched at the transcription termination sites (TTS) of highly expressed, convergent genes and that replication stress markers are also enriched at these regions suggesting a mechanism of replication-transcription conflicts that is resolved by topoisomerases (TOP) in normal cells considering that double strand breaks form in TOP1-depleted cells [96]. Evidence also showed that the mutational signature ID4 referenced in somatic mutation cancer database COSMIC is similar to the signature they extracted which is associated with the defective activity of TOP1 at sites where ribonucleotides were mis-incorporated [168] which consistent with the fact that ribonucleotides that get embedded in the DNA sequence can be a source of replicative stress [1].

Considering the above strong evidence suggests that the replication program and the replication stress play a significant role in shaping the mutational landscape of a cancer genome. Hence, we hypothesize that there might be mutational signatures specific to replication stress or more specifically linked to TRC. Furthermore, there could be an association between the rates at which certain mutational processes operates and replication stress occurrence. To that end, future study can be to aim at searching for mutational signatures associated with replication stress. More specifically, combining these data with our established TRC pipeline, it is possible to verify whether there are different mutational signatures at TSS and TTS within potential hotspot genes and if there is a mutational asymmetry in this context.

4. Study of direct detecting TRCs in genome-wide even in single-cell

Studying transcription and replication during the same experiment is highly challenging and no well-developed technique can identify TRCs. Nevertheless, there is a recent technology called transcription-replication immunoprecipitation on nascent DNA sequencing (TRIPn-seq) that tries to fill this gap [169]. This technique is based on the immunoprecipitation of the RNA polymerase 2 phosphorylated at serine 5 (RNAP2s5), phosphorylation present during elongation, followed by a second immunoprecipitation of previously BrdU-labeled nascent DNA. This technique is very promising to further investigate TRC, but this protocol still needs to be tuned and the produce data have to be validated.

Other strategies to study TRC could be based on cell-cycle RNA-seq or more specifically the S phase GRO-seq which would allow to map the transcription machinery within S phase and compare its position with OK-seq data or any other techniques containing fork orientation information. However, limiting to S phase does not mean that the replication and transcription machineries are simultaneous present on the same allele in the same cell.

Therefore, one technical possibility is to label these two types of enzymes respectively, or theoretically tag the relevant proteins to map the genome-wide polymerases simultaneously, which is able to find the colocalization of the transcription and replication in spatial and also in temporal dimension even possibly extending to every single cell. The protein-protein colocalization could be inspired from some recent techniques such as proximity ligation assay (PLA) [170] using specific antibodies to replisome components (e.g., PCNA, ORC, MCM or DNA polymerases) and RNAPII following by fluorescent hybridization with PCR which could give a first visualization of the colocalization spots in fluorescence microscopy; CUT&RUN [171] and CUT&Tag [172] techniques can be considered for establishing the genome-wide transcription-replication interaction profiles using double ChIP strategy which allows to bind the protein of interest in situ by a specific antibody, then tethers by A-fused transposase Tn5. CUT&Tag technique has been successfully profiling RNAPII and other transcription factors only low start materials needed and also in single-cell level by loading single-cell barcodes on Tn5 transposase [172], which make this strategy technically feasible in the future.

Following another approach, we can take advantage of a newly developed techniques such as ORM. This would allow us to detect the initiation events and fork progression in single-molecule level [61] using three fluorescent color which the green fluorophore tag specific motif sequences and used for mapping to reference genome while red one used to label the ongoing

replication origins. It may possibly apply to map the ongoing RNAPII to get the transcription progression in single-molecule level or even further optimize the Bionano technique for adding one more fluorophore tag in the proximity of RNAP that could achieve to detect both replication and transcription process.

5. Study of R-loop detection in fork stalling and restart

Another challenge along the previous TRC research is when and where 'toxic' R-loops actually formed while encountering replication progression. As mentioned in *Chapter 1-3.3.1*, the real mechanism implicated in R-loop related replication fork stalling and restart is still in debate. There is no consensus on how to explain the way TRCs, RNA polymerase and R-loops interfere block the fork progression.

The most well-known model suggests that RNAP and/or R-loop interfere confrontationally with replication machinery by directly blocking the fork progression, while R-loops are supposed to be formed ahead of forks and RNase H is recruited to solve R-loops to prevent fork stalling [86,89]. In our own research, we also observed a replication speed rescue with overexpression of RNase H in both control and shTOP1 cells [96]. In this context, R-loop can be considered as a cause of fork stalling. However, some more recent studies also suggest that R-loops could also accumulate behind the forks as a consequence of fork stalling and be involved in the fork restart processes, such as fork reversal or resection [92]. Hence R-loops could be favored by the presence of ssDNA gaps and persist after fork pass by. In this case, RNase H would be required to promote fork restart [86] and R-loop accumulation is rather as a consequence of fork stalling and restart.

It has to be noticed that CMG complex travels on the leading strand while R-loops form abundantly on the lagging strand. Therefore, trying to detect the strand-specific R-loops could be more informative. Nevertheless, a recent paper compared the most classic RNA:DNA hybrid sequencing techniques based on S9.6 antibody approach with a novel dRNase H1 approach, which is using defective RNase H1 (catalytically inactive form of RNase H1) that is able to recognize but not to process RNA:DNA hybrids [173]. Surprisingly the two approaches have generally disparate hybrid mapping results and the comparison of specificity/affinity between dRNase H1 and S 9.6 to RNA:DNA hybrid is still unclear [173]. As the point of views mentioned above, RNA:DNA hybrid detecting in single-cell/single-molecule could be the future direction to identify the true role of R-loops in fork stalling process and also in the TRC regulation.

Chapter 5 References

1. Zeman, M. K. & Cimprich, K. A. Causes and consequences of replication stress. *Nat. Cell Biol.* **16**, 2–9 (2014).
2. Gnan, S., Liu, Y., Spagnuolo, M. & Chen, C.-L. The impact of transcription-mediated replication stress on genome instability and human disease. *Genome Instab. Dis.* 1–28 (2020). doi:10.1007/s42764-020-00021-y
3. Hamperl, S., Bocek, M. J., Saldivar, J. C., Swigut, T. & Cimprich, K. A. Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses. *Cell* **170**, 774-786.e19 (2017).
4. Niehrs, C. & Luke, B. Regulatory R-loops as facilitators of gene expression and genome stability. *Nat. Rev. Mol. Cell Biol.* **21**, 167–178 (2020).
5. García-Muse, T. & Aguilera, A. Transcription–replication conflicts: how they occur and how they are resolved. *Nat. Rev. Mol. Cell Biol.* **17**, 553–563 (2016).
6. Daigaku, Y. *et al.* A global profile of replicative polymerase usage. *Nat. Struct. Mol. Biol.* (2015). doi:10.1038/nsmb.2962
7. Hennion, M. *et al.* FORK-seq: Replication landscape of the Saccharomyces cerevisiae genome by nanopore sequencing. *Genome Biol.* **21**, 1–25 (2020).
8. Sriramachandran, A. M. *et al.* Genome-wide Nucleotide-Resolution Mapping of DNA Replication Patterns, Single-Strand Breaks, and Lesions by GLOE-Seq. *Mol. Cell* **78**, 975-985.e7 (2020).
9. Petryk, N. *et al.* MCM2 promotes symmetric inheritance of modified histones during DNA replication. *Science (80-.).* **361**, 1389–1392 (2018).
10. Li, Z. *et al.* DNA polymerase α interacts with H3-H4 and facilitates the transfer of parental histones to lagging strands. *Sci. Adv.* **6**, (2020).
11. Kara, N., Krueger, F., Rugg-Gunn, P. & Houseley, J. *Genome-wide analysis of DNA replication and DNA double-strand breaks using TrAEL-seq*. *PLOS Biology* **19**, (2021).
12. Petryk, N. *et al.* Replication landscape of the human genome. *Nat. Commun.* **7**, 10208 (2016).
13. Barnum, K. J. & O'Connell, M. J. Cell cycle regulation by checkpoints. *Methods Mol. Biol.* **1170**, 29–40 (2014).
14. GM, C. *The Cell: A Molecular Approach.* (2000).

15. Giacinti, C. & Giordano, A. RB and cell cycle progression. *Oncogene* **25**, 5220–5227 (2006).
16. Barnaba, N. & LaRocque, J. R. Targeting cell cycle regulation via the G2-M checkpoint for synthetic lethality in melanoma. *Cell Cycle* **20**, 1041–1051 (2021).
17. Stark, G. R. & Taylor, W. R. Analyzing the G2/M checkpoint. *Methods Mol. Biol.* **280**, 51–82 (2004).
18. Cayrou, C., Grégoire, D., Coulombe, P., Danis, E. & Méchali, M. Genome-scale identification of active DNA replication origins. *Methods* **57**, 158–164 (2012).
19. Kang, S., Warner, M. D. & Bell, S. P. Multiple Functions for Mcm2-7 ATPase Motifs during Replication Initiation. *Mol. Cell* **55**, 655–665 (2014).
20. Lin, Y. C. & Prasanth, S. G. Replication initiation: Implications in genome integrity. *DNA Repair (Amst).* **103**, 103131 (2021).
21. Kanke, M., Kodama, Y., Takahashi, T. S., Nakagawa, T. & Masukata, H. Mcm10 plays an essential role in origin DNA unwinding after loading of the CMG components. *EMBO J.* (2012). doi:10.1038/emboj.2012.68
22. Yeeles, J. T. P., Deegan, T. D., Janska, A., Early, A. & Diffley, J. F. X. Regulated eukaryotic DNA replication origin firing with purified proteins. *Nature* **519**, 431–435 (2015).
23. Costa, A. & Diffley, J. F. X. The initiation step of eukaryotic DNA replication. *Annu. Rev. Biochem.* (2022). doi:10.1007/978-90-481-3471-7_5
24. Zegerman, P. Evolutionary conservation of the CDK targets in eukaryotic DNA replication initiation. *Chromosoma* **124**, 309–321 (2015).
25. Kang, S., Kang, M. S., Ryu, E. & Myung, K. Eukaryotic DNA replication: Orchestrated action of multi-subunit protein complexes. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **809**, 58–69 (2018).
26. Pellegrini, L. & Costa, A. New Insights into the Mechanism of DNA Duplication by the Eukaryotic Replisome. *Trends Biochem. Sci.* **41**, 859–871 (2016).
27. Branzei, D. & Foiani, M. Maintaining genome stability at the replication fork. *Nat. Rev. Mol. Cell Biol.* **11**, 208–219 (2010).
28. Ge, X. Q., Jackson, D. A. & Blow, J. J. Dormant origins licensed by excess Mcm2-7 are required for human cells to survive replicative stress. *Genes Dev.* (2007). doi:10.1101/gad.457807
29. MacNeill, S. A. DNA replication: Partners in the Okazaki two-step. *Curr. Biol.* **11**, 842–844 (2001).

30. Dewar, J. M. & Walter, J. C. Mechanisms of DNA replication termination. *Nat. Rev. Mol. Cell Biol.* **18**, 507–516 (2017).
31. Donley, N. & Thayer, M. J. DNA replication timing, genome stability and cancer. Late and/or delayed DNA replication timing is associated with increased genomic instability. *Semin. Cancer Biol.* **23**, 80–89 (2013).
32. Dimitrova, D. S. & Gilbert, D. M. The spatial position and replication timing of chromosomal domains are both established in early G1 phase. *Mol. Cell* **4**, 983–993 (1999).
33. Koren, A. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* **91**, 1033–1040 (2012).
34. Ryba, T., Battaglia, D., Pope, B. D., Hiratani, I. & Gilbert, D. M. Genome-Scale Analysis of Replication Timing: from Bench to Bioinformatics. *Nat. Protoc.* (2011). doi:10.1038/nprot.2011.328.Genome-Scale
35. Gilbert, D. M. Evaluating genome-scale approaches to eukaryotic DNA replication. *Nat. Publ. Gr.* **11**, 673–684 (2010).
36. Claire Marchal, Takayo Sasaki, Daniel Vera, Korey Wilson, Jiao Sima, Juan Carlos Rivera-Mulia, Claudia Trevilla-García, Coralin Nogues, Ebtesam Nafie, and D. M. G. Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq. *Nat. Protoc.* **13**, 819–839 (2018).
37. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci.* **107**, 139–144 (2010).
38. Chun-Long Chen, Aurelien Rappailles, Lauranne Duquenne, M. H., Guillaume Guilbaud, Laurent Farinelli, Benjamin Audit, Y. d'Aubenton-C. & Alain Arneodo, O. H. and C. T. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res.* (2010).
39. Desprat, R. *et al.* Predictable dynamic program of timing of DNA replication in human cells Predictable dynamic program of timing of DNA replication in human cells. *Genome Res.* 2288–2299 (2009). doi:10.1101/gr.094060.109
40. Dileep, V. & Gilbert, D. M. Single-cell replication profiling to measure stochastic variation in mammalian replication timing. *Nat. Commun.* **9**, (2018).
41. Zhao, P. A., Sasaki, T. & Gilbert, D. M. High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome Biol.* **21**, 1–20 (2020).
42. Allemand, J. F., Bensimon, D., Jullien, L., Bensimon, A. & Croquette, V. pH-

dependent specific binding and combing of DNA. *Biophys. J.* **73**, 2064–2070 (1997).

43. Bianco, J. N. *et al.* Analysis of DNA replication profiles in budding yeast and mammalian cells using DNA combing. *Methods* **57**, 149–157 (2012).
44. Fu, H. & Aladjem, M. I. DNA replication profiling by molecular combing on single DNA fibers. *STAR Protoc.* **3**, 101290 (2022).
45. Kaykov, A., Taillefumier, T., Bensimon, A. & Nurse, P. Molecular Combing of Single DNA Molecules on the 10 Megabase Scale. *Sci. Rep.* **6**, 80–84 (2016).
46. Müller, C. A. *et al.* Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads. *Nat. Methods* (2019). doi:10.1038/s41592-019-0394-y
47. Theulot, B. *et al.* Genome-wide mapping of individual replication fork velocities using nanopore sequencing. *Nat. Commun.* **13**, 1–14 (2022).
48. Giacca, M., Pelizon, C. & Falaschi, A. Mapping replication origins by quantifying relative abundance of nascent DNA strands using competitive polymerase chain reaction. *Methods A Companion to Methods Enzymol.* **13**, 301–312 (1997).
49. Prioleau, M.-N., Gendron, M.-C. & Hyrien, O. Replication of the Chicken β-Globin Locus: Early-Firing Origins at the 5′ HS4 Insulator and the ρ- and β A -Globin Genes Show Opposite Epigenetic Modifications . *Mol. Cell. Biol.* **23**, 3536–3549 (2003).
50. Picard, F. *et al.* The Spatiotemporal Program of DNA Replication Is Associated with Specific Combinations of Chromatin Marks in Human Cells. *PLoS Genet.* **10**, (2014).
51. Haiqing Fu *et al.* Mapping Replication Origin Sequences in Eukaryotic Chromosomes. *curr protoc cell biol* (2015). doi:10.1002/0471143030.cb2220s65.Mapping
52. Mesner, L. D. & Hamlin, J. L. Isolation of restriction fragments containing origins of replication from complex genomes. *Methods Mol. Biol.* **1300**, 279–292 (2009).
53. Mesner, L. D., Crawford, E. L. & Hamlin, J. L. Isolating apparently pure libraries of replication origins from complex genomes. *Mol. Cell* **21**, 719–726 (2006).
54. Smith, D. J. & Whitehouse, I. Intrinsic coupling of lagging-strand synthesis to chromatin assembly. *Nature* **483**, 434–438 (2012).
55. Tyteca, S., Vandromme, M., Legube, G., Chevillard-Briet, M. & Trouche, D. Tip60 and p400 are both required for UV-induced apoptosis but play antagonistic roles in cell cycle progression. *EMBO J.* **25**, 1680–1689 (2006).
56. Sugimoto, N., Maehara, K., Yoshida, K., Ohkawa, Y. & Fujita, M. Genome-wide analysis of the spatiotemporal regulation of firing and dormant replication origins in human cells. *Nucleic Acids Res.* **46**, 6683–6696 (2018).
57. Kirstein, N. *et al.* Human ORC/MCM density is low in active genes and correlates with

replication time but does not delimit initiation zones. *Elife* **10**, 1–30 (2021).

58. Hua, H. & Kearsey, S. E. Monitoring DNA replication in fission yeast by incorporation of 5-ethynyl-2′-deoxyuridine. *Nucleic Acids Res.* **39**, (2011).
59. Macheret, M. & Halazonetis, T. D. Monitoring early S-phase origin firing and replication fork movement by sequencing nascent DNA from synchronized cells. *Nat. Protoc.* **14**, 51–67 (2019).
60. Langley, A. R., Gräf, S., Smith, J. C. & Krude, T. Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). *Nucleic Acids Res.* **44**, 10230–10247 (2016).
61. Wang, W. *et al.* Genome-wide mapping of human DNA replication by optical replication mapping supports a stochastic model of eukaryotic replication. *Mol. Cell* **81**, 2975-2988.e6 (2021).
62. Touchon, M. *et al.* Replication-associated strand asymmetries in mammalian genomes: Toward detection of replication origins. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 9836–9841 (2005).
63. Huvet, M. *et al.* Human gene organization driven by the coordination of replication and transcription. *Genome Res.* **17**, 1278–85 (2007).
64. Chen, C. L. *et al.* Replication-associated mutational asymmetry in the human genome. *Mol. Biol. Evol.* **28**, 2327–2337 (2011).
65. Hyrien, O. *et al.* From simple bacterial and archaeal replicons to replication N/U-domains. *J. Mol. Biol.* **425**, 4673–4689 (2013).
66. Clausen, A. R. Tracking replication enzymology in vivo by genome-wide mapping of ribonucleotide incorporation. *Nat Struct Mol Biol* (2015). doi:10.1038/nsmb.2957.Tracking
67. Koyanagi, E. *et al.* Global landscape of replicative DNA polymerase usage in the human genome. *bioRxiv* 2021.11.14.468503 (2021).
68. Yu, C. *et al.* Strand-Specific Analysis Shows Protein Binding at Replication Forks and PCNA Unloading from Lagging Strands when Forks Stall. *Mol. Cell* **56**, 551–563 (2014).
69. Atkinson, T. J. & Halfon, M. S. Regulation of gene expression in the genomic context. *Comput. Struct. Biotechnol. J.* **9**, e201401001 (2014).
70. Proudfoot, N. J. Transcriptional translation in mammals: Stopping de RNA polymerase II juggernaut. *Science (80-.).* **352**, (2016).
71. Dominguez, D. *et al.* An extensive program of periodic alternative splicing linked to

cell cycle progression. *Elife* **5**, 1–19 (2016).

72. Macheret, M. & Halazonetis, T. D. Intragenic origins due to short G1 phases underlie oncogene-induced DNA replication stress. *Nature* **555**, 112–116 (2018).
73. Hamperl, S. & Cimprich, K. A. Conflict Resolution in the Genome: How Transcription and Replication Make It Work. *Cell* **167**, 1455–1467 (2016).
74. Kunst, F. *et al.* The complete genome sequence of the Gram-positive bacterium Bacillus subtilis. *Nature* **390**, 249–256 (1997).
75. Yu-Hung Chen, Sarah Keegan, Malik Kahli, Peter Tonzi, David Fenyö, Tony T. Huang, and D. J. S. Transcription shapes DNA replication initiation and termination in human cells. *Nat. Struct. Mol. Biol.* **26**, 67–77 (2019).
76. Srivatsan, A., Tehranchi, A., MacAlpine, D. M. & Wang, J. D. Co-orientation of replication and transcription preserves genome integrity. *PLoS Genet.* **6**, (2010).
77. Thomas, M., White, R. L. & Davis, R. W. Hybridization of RNA to double-stranded DNA: formation of R-loops. *Proc. Natl. Acad. Sci. U. S. A.* **73**, 2294–8 (1976).
78. Groh, M. & Gromak, N. Out of Balance: R-loops in Human Disease. *PLoS Genet.* **10**, (2014).
79. Yu, K., Chedin, F., Hsieh, C. L., Wilson, T. E. & Lieber, M. R. R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. *Nat. Immunol.* **4**, 442–451 (2003).
80. Sanz, L. A. *et al.* Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Mol. Cell* **63**, 167–178 (2016).
81. Ginno, P. A., Lott, P. L., Christensen, H. C., Korf, I. & Chédin, F. R-Loop Formation Is a Distinctive Characteristic of Unmethylated Human CpG Island Promoters. *Mol. Cell* **45**, 814–825 (2012).
82. Skourti-Stathaki, K., Proudfoot, N. J. & Gromak, N. Human Senataxin Resolves RNA/DNA Hybrids Formed at Transcriptional Pause Sites to Promote Xrn2-Dependent Termination. *Mol. Cell* **42**, 794–805 (2011).
83. Holt, I. J. Survey and summary: The mitochondrial R-loop. *Nucleic Acids Res.* **47**, 5480–5489 (2019).
84. Kabeche, L., Nguyen, H. D., Buisson, R. & Zou, L. A mitosis-specific and R loop–driven ATR pathway promotes faithful chromosome segregation. *Science (80-.).* **359**, 108 LP – 114 (2018).
85. Skourti-Stathaki, K. & Proudfoot, N. J. A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. *Genes Dev* **28 SRC**-,

1384–1396 (2014).

86. Kemiha, S., Poli, J., Lin, Y. L., Lengronne, A. & Pasero, P. Toxic R-loops: Cause or consequence of replication stress? *DNA Repair (Amst).* **107**, (2021).
87. Crossley, M. P., Bocek, M. & Cimprich, K. A. R-Loops as Cellular Regulators and Genomic Threats. *Mol. Cell* **73**, 398–411 (2019).
88. Stork, C. T. *et al.* Co-transcriptional R-loops are the main cause of estrogen-induced DNA damage. *Elife* **5**, 1–21 (2016).
89. Brickner, J. R., Garzon, J. L. & Cimprich, K. A. Walking a tightrope : The complex balancing act of R-loops in genome stability. *Mol. Cell* (2022). doi:10.1016/j.molcel.2022.04.014
90. Zhang, C. *et al.* METTL3 and N6-Methyladenosine Promote Homologous Recombination-Mediated Repair of DSBs by Modulating DNA-RNA Hybrid Accumulation. *Mol. Cell* **79**, 425-442.e7 (2020).
91. Maffia, A., Ranise, C. & Sabbioneda, S. From R-loops to G-quadruplexes: Emerging new threats for the replication fork. *Int. J. Mol. Sci.* **21**, (2020).
92. Barroso, S. *et al.* The DNA damage response acts as a safeguard against harmful DNA–RNA hybrids of different origins. *EMBO Rep.* **20**, (2019).
93. Blackford, A. N. & Jackson, S. P. ATM, ATR, and DNA-PK: The Trinity at the Heart of the DNA Damage Response. *Mol. Cell* **66**, 801–817 (2017).
94. Zou, L. & Elledge, S. J. Sensing DNA damage through ATRIP recognition of RPA-ssDNA complexes. *Science (80-.).* **300**, 1542–1548 (2003).
95. Liu, Y., Wu, X., D'aubenton-Carafa, Y., Thermes, C. & Chen, C.-L. OKseqHMM: a genome-wide replication fork directionality analysis toolkit. *bioRxiv* (2022). doi:10.1101/2022.01.12.476022
96. Promonet, A. *et al.* Topoisomerase 1 prevents replication stress at R-loop-enriched transcription termination sites. *Nat. Commun.* **11**, 3940 (2020).
97. Liu, Y., Lin, Y. L., Pasero, P. & Chen, C. L. Topoisomerase I prevents transcription-replication conflicts at transcription termination sites. *Mol. Cell. Oncol.* **8**, (2021).
98. Merrikh, H. Spatial and Temporal Control of Evolution through Replication–Transcription Conflicts. *Trends Microbiol.* **25**, 515–521 (2017).
99. Chen, C. L. *et al.* Replication-associated mutational asymmetry in the human genome. *Mol. Biol. Evol.* **28**, 2327–2337 (2011).
100. Haradhvala, N. J. *et al.* Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **164**, 538–549 (2016).

101. Cortez, L. M. *et al. APOBEC3A is a prominent cytidine deaminase in breast cancer*. *PLoS Genetics* **15**, (2019).
102. Hoopes, J. I. *et al.* APOBEC3A and APOBEC3B Preferentially Deaminate the Lagging Strand Template during DNA Replication. *Cell Rep.* **14**, 1273–1282 (2016).
103. Shi, M. J. *et al.* APOBEC-mediated Mutagenesis as a Likely Cause of FGFR3 S249C Mutation Over-representation in Bladder Cancer. *Eur. Urol.* **76**, 9–13 (2019).
104. Shi, M. J. *et al.* Identification of new driver and passenger mutations within APOBEC-induced hotspot mutations in bladder cancer. *Genome Med.* **12**, 1–20 (2020).
105. Mas-Ponte, D. & Supek, F. DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers. *Nat. Genet.* **52**, 958–968 (2020).
106. Shi, Y. *et al.* Reduced Expression of METTL3 Promotes Metastasis of Triple-Negative Breast Cancer by m6A Methylation-Mediated COL3A1 Up-Regulation. *Front. Oncol.* **10**, 1–15 (2020).
107. Wang, H., Xu, B. & Shi, J. N6-methyladenosine METTL3 promotes the breast cancer progression via targeting Bcl-2. *Gene* **722**, 144076 (2020).
108. Huang, H., Weng, H. & Chen, J. m6A Modification in Coding and Non-coding RNAs: Roles and Therapeutic Implications in Cancer. *Cancer Cell* **37**, 270–288 (2020).
109. Petryk, N. *et al.* Replication landscape of the human genome. *Nat. Commun.* **7**, 10208 (2016).
110. McGuffee, S. R., Smith, D. J. & Whitehouse, I. Quantitative, Genome-Wide Analysis of Eukaryotic Replication Initiation and Termination. *Mol. Cell* **50**, 123–135 (2013).
111. Wu, X. *et al.* Developmental and cancer-associated plasticity of DNA replication preferentially targets GC-poor, lowly expressed and late-replicating regions. *Nucleic Acids Res.* 1–16 (2018). doi:10.1093/nar/gky797
112. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160-5 (2016).
113. Siow, C. C., Nieduszynska, S. R., Müller, C. A. & Nieduszynski, C. A. OriDB, the DNA replication origin database updated and extended. *Nucleic Acids Res.* **40**, 682–686 (2012).
114. Gnan, S. *et al.* Kronos scRT: a uniform framework for single-cell replication timing analysis. *Nat. Commun.* **13**, s41467-022-30043–x (2022).
115. Aguilera, A. & García-Muse, T. Causes of Genome Instability. *Annu. Rev. Genet.* **47**, 1–32 (2013).
116. Saldivar, J. C. *et al.* An intrinsic S/G 2 checkpoint enforced by ATR. *Science (80-.).*

361, 806–810 (2018).

117. El Hage, A., French, S. L., Beyer, A. L. & Tollervey, D. Loss of Topoisomerase I leads to R-loop-mediated transcriptional blocks during ribosomal RNA synthesis. *Genes Dev.* (2010). doi:10.1101/gad.573310
118. Iacovoni, J. S. *et al.* High-resolution profiling of γh2AX around DNA double strand breaks in the mammalian genome. *EMBO J.* **29**, 1446–1457 (2010).
119. Biernacka, A. *et al.* i-BLESS is an ultra-sensitive method for detection of DNA double-strand breaks. *Commun. Biol.* **1**, (2018).
120. Andersson, R. *et al.* Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat. Commun.* **5**, (2014).
121. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
122. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
123. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
124. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, (2008).
125. Quinlan, A. R. *BEDTools: The Swiss-Army tool for genome feature analysis*. *Current Protocols in Bioinformatics* **2014**, (2014).
126. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, (2004).
127. Manzo, S. G. *et al.* DNA Topoisomerase I differentially modulates R-loops across the human genome. *Genome Biol.* **19**, 1–18 (2018).
128. Tuduri, S. *et al.* Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription. *Nat. Cell Biol.* **11**, 1315–1324 (2009).
129. Li, X. & Manley, J. L. Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. *Cell* **122**, 365–378 (2005).
130. Drolet, M. *et al.* Overexpression of RNase H partially complements the growth defect of an Escherichia coli ΔtopA mutant: R-loop formation is a major problem in the absence of DNA topoisomerase I. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 3526–3530 (1995).
131. García-Rubio, M., Huertas, P., González-Barrera, S. & Aguilera, A. Recombinogenic Effects of DNA-Damaging Agents Are Synergistically Increased by Transcription in

Saccharomyces cerevisiae: New Insights into Transcription-Associated Recombination. *Genetics* **165**, 457–466 (2003).

132. Chappidi, N. *et al.* Fork Cleavage-Religation Cycle and Active Transcription Mediate Replication Restart after Fork Stalling at Co-transcriptional R-Loops. *Mol. Cell* **77**, 528-541.e8 (2020).
133. Marnef, A., Cohen, S. & Legube, G. Transcription-Coupled DNA Double-Strand Break Repair: Active Genes Need Special Care. *J. Mol. Biol.* **429**, 1277–1288 (2017).
134. Canela, A. *et al.* Topoisomerase II-Induced Chromosome Breakage and Translocation Is Determined by Chromosome Architecture and Transcriptional Activity. *Mol. Cell* **75**, 252-266.e8 (2019).
135. Seiler, J. A., Conti, C., Syed, A., Aladjem, M. I. & Pommier, Y. The Intra-S-Phase Checkpoint Affects both DNA Replication Initiation and Elongation: Single-Cell and -DNA Fiber Analyses. *Mol. Cell. Biol.* **27**, 5806–5818 (2007).
136. Mutreja, K. *et al.* ATR-Mediated Global Fork Slowing and Reversal Assist Fork Traverse and Prevent Chromosomal Breakage at DNA Interstrand Cross-Links. *Cell Rep.* **24**, 2629-2642.e5 (2018).
137. Bacal, J. *et al.* Mrc1 and Rad9 cooperate to regulate initiation and elongation of DNA replication in response to DNA damage. *EMBO J.* **37**, 1–18 (2018).
138. Osmundson, J. S., Kumar, J., Yeung, R. & Smith, D. J. Pif1-family helicases cooperate to suppress widespread replication fork arrest at tRNA genes. *Nat Struct Mol Biol* **24**, 162–170 (2017).
139. Cohen, S. *et al.* Senataxin resolves RNA:DNA hybrids forming at DNA double-strand breaks to prevent translocations. *Nat. Commun.* **9**, (2018).
140. Tran, P. L. T. *et al.* PIF1 family DNA helicases suppress R-loop mediated genome instability at tRNA genes. *Nat. Commun.* **8**, (2017).
141. Nguyen, V. C. *et al.* Replication stress checkpoint signaling controls tRNA gene transcription. *Nat. Struct. Mol. Biol.* **17**, 976–981 (2010).
142. Poli, J. *et al.* Mec1, INO80, and the PAF1 complex cooperate to limit transcription replication conflicts through RNAPII removal during replication stress. *Genes Dev.* **30**, 337–354 (2016).
143. Lafon, A. *et al.* INO80 Chromatin Remodeler Facilitates Release of RNA Polymerase II from Chromatin for Ubiquitin-Mediated Proteasomal Degradation. *Mol. Cell* **60**, 784–796 (2015).
144. Rodriguez, J. *et al.* Intrinsic Dynamics of a Human Gene Reveal the Basis of

Expression Heterogeneity. *Cell* **176**, 213-226.e18 (2019).

145. Teloni, F. *et al.* Efficient Pre-mRNA Cleavage Prevents Replication-Stress-Associated Genome Instability. *Mol. Cell* **73**, 670-683.e12 (2019).
146. Lang, K. S. *et al.* Replication-Transcription Conflicts Generate R-Loops that Orchestrate Bacterial Stress Survival and Pathogenesis. *Cell* **263**, 219–227 (2017).
147. Costantino, L. & Koshland, D. Genome-wide Map of R-Loop-Induced Damage Reveals How a Subset of R-Loops Contributes to Genomic Instability. *Mol. Cell* **71**, 487-497.e3 (2018).
148. Harbeck, N. *et al.* Breast cancer. *Nat. Rev. Dis. Prim.* **5**, 66 (2019).
149. Samavat, H. & Kurzer, M. S. Estrogen Metabolism and Breast Cancer. *Cancer Lett.* (2015). doi:10.1016/j.canlet.2014.04.018.Estrogen
150. Williamson, L. M. & Lees-Miller, S. P. Estrogen receptor α-mediated transcription induces cell cycle-dependent DNA double-strand breaks. *Carcinogenesis* **32**, 279–285 (2011).
151. Abakir, A. *et al.* N 6-methyladenosine regulates the stability of RNA:DNA hybrids in human cells. *Nat. Genet.* **52**, 48–55 (2020).
152. Marnef, A. & Legube, G. m6A RNA modification as a new player in R-loop regulation. *Nat. Genet.* **52**, 27–28 (2020).
153. Yang, X. *et al.* m6A promotes R-loop formation to facilitate transcription termination. *Cell Res.* **29**, 1035–1038 (2019).
154. Moreira, M. C. *et al.* Senataxin, the ortholog of a yeast RNA helicase, is mutant in ataxia-ocular apraxia 2. *Nat. Genet.* **36**, 225–227 (2004).
155. Steinmetz, E. J. *et al.* Genome-Wide Distribution of Yeast RNA Polymerase II and Its Control by Sen1 Helicase. *Mol. Cell* **24**, 735–746 (2006).
156. Lim, Y. W., Sanz, L. A., Xu, X., Hartono, S. R. & Chédin, F. Genome-wide DNA hypomethylation and RNA:DNA hybrid accumulation in Aicardi-Goutières syndrome. *Elife* **4**,
157. Rice, G. I. Mutations involved in Aicardi-Goutieres syndrome implicate SAMHD1 as regulator of the innate immune response. *Nat. Genet.* **23**, 1–7 (2009).
158. Cristini, A. *et al.* RNase H2, mutated in Aicardi-Goutières syndrome, resolves co-transcriptional R-loops to prevent DNA breaks and inflammation. *Nat. Commun.* **13**, 1–14 (2022).
159. Groh, M., Lufino, M. M. P., Wade-Martins, R. & Gromak, N. R-loops associated with triplet repeat expansions promote gene silencing in Friedreich ataxia and fragile X

syndrome. *PLoS Genet.* **10**, e1004318 (2014).

160. Stirling, P. C. *et al.* R-loop-mediated genome instability in mRNA cleavage and polyadenylation mutants. *Genes Dev.* **26**, 163–175 (2012).
161. Gorthi, A. EWS–FLI1 increases transcription to cause R-loops and block BRCA1 repair in Ewing sarcoma. *Nature* **176**, 139–148 (2018).
162. Li, Z. *et al.* A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 8164–8169 (2003).
163. Shou, Y. *et al.* Diverse karyotypic abnormalities of the c-myc locus associated with c-myc dysregulation and tumor progression in multiple myeloma. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 228–233 (2000).
164. Küppers, R. & Dalla-Favera, R. Mechanisms of chromosomal translocations in B cell lymphomas. *Oncogene* **20**, 5580–5594 (2001).
165. Li, Y. *et al.* R-loops coordinate with SOX2 in regulating reprogramming to pluripotency. *Sci. Adv.* **6**, (2020).
166. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
167. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
168. Reijns, M. A. M. *et al.* Signatures of TOP1 transcription-associated mutagenesis in cancer and germline. *Nature* (2022). doi:10.1038/s41586-022-04403-y
169. St Germain, C. P. *et al.* Genomic patterns of transcription-replication interactions in mouse primary B cells. *Nucleic Acids Res.* 2021.04.13.439211 (2022).
170. Alam, M. S. Proximity Ligation Assay (PLA). *Curr. Protoc. Immunol.* **176**, 139–148 (2018).
171. Skene, P. J., Henikoff, J. G. & Henikoff, S. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nat. Protoc.* **13**, 1006–1019 (2018).
172. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1–10 (2019).
173. Hartono, S. R., Sanz, L. A. & Vanoosthuyse, V. Best practices for the visualization , mapping , and manipulation of R-loops. 1–13 (2021). doi:10.15252/embj.2020106394

Chapter 6 Annex

Annex 1

Genome-wide measurement of DNA replication fork directionality and quantification of DNA replication initiation and termination with Okazaki fragment sequencing

Xia Wu[#], Yaqun Liu[#], Yves d'Aubenton-Carafa, Claude Thermes, Olivier Hyrien*, Chun-Long Chen*, Nataliya Petryk*.

(Accepted by *Nature Protocols*)

Genome-wide measurement of DNA replication fork directionality and quantification of DNA replication initiation and termination with Okazaki fragment sequencing

Xia Wu[1,#]; Yaqun Liu[2,#], Yves d'Aubenton-Carafa[3], Claude Thermes[3], Olivier Hyrien[4,*], Chun-Long Chen[2,*], Nataliya Petryk[5,*].

1. *Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, Guangdong, 510080, China.*
2. *Institut Curie, Université PSL, Sorbonne Université, CNRS UMR3244, Dynamics of Genetic Information, 26 rue d'Ulm, Paris, 75005, France.*
3. *Institute for Integrative Biology of the Cell (I2BC), Université Paris-Saclay, CEA, CNRS, Gif-sur-Yvette, 91198, France.*
4. *Institut de Biologie de l'École Normale Supérieure (IBENS), École Normale Supérieure, CNRS, Inserm, Université PSL, 46 rue d'Ulm, 75005 Paris, France.*
5. *Epigenetics & cell fate CNRS UMR7216 University of Paris, 35 rue Hélène Brion, Paris, 75013, France.*

[#] These authors contributed equally

[*]Correspondence to OH, CLC, and NP: olivier.hyrien@bio.ens.psl.eu; chunlong.chen@curie.fr; nataliya.petryk@cnrs.fr

ABSTRACT

Studying the dynamics of genome replication in mammalian cells has been historically challenging. To reveal the location of replication initiation and termination in the human genome, we developed OK-seq, a quantitative approach based on the isolation and strand-specific sequencing of Okazaki fragments, the lagging strand replication intermediates. OK-seq quantitates the proportion of leftward- and rightward-oriented forks at every genomic locus and reveals the location and efficiency of replication initiations and terminations. Here we provide the detailed experimental procedures for performing OK-seq in unperturbed cultured human cells and budding yeasts and the bioinformatics pipelines for data processing and computation of replication fork directionality (RFD). Furthermore, we present the analytical approach based on a hidden Markov model (HMM), which allows automated

detection of ascending, descending, and flat RFD segments revealing the zones of replication initiation, termination, and unidirectional fork movement across the entire genome. These tools are essential for accurate interpretation of human and yeast replication programs. Besides revealing the genome replication program in fine detail, OK-seq has been instrumental in numerous studies unravelling mechanisms of genome stability, epigenome maintenance, and genome evolution.

INTRODUCTION

DNA fibre autoradiographic studies of mammalian cells showed long ago that eukaryotic DNA replication origins are spaced at 20-400 kb intervals and fire at different times in S phase [1]. However, mapping origins in metazoan cells has been historically challenging, due to the lack of workable genetic assays and the difficulties in purifying sufficient amounts of intact DNA replication initiation intermediates (for reviews [2-4]).

In the pre-genomic era, early studies of the highly amplified CHO DHFR locus identified a few specific initiation sites downstream of the DHFR gene. However, more extensive studies demonstrated that replication could initiate at any of a large number of sites over a broad (55 kb) zone downstream of the gene, at a global rate lower than one initiation event per cell cycle, even in unamplified cells [2]. Depending on the technique(s) used to purify initiation intermediates from cell populations, site-specific or dispersed initiation was also reported at a few other model loci [3]. Direct visualization of replication fork progression at the single DNA molecule level using DNA combing [5] or SMARD [6] in general revealed broad (3-100 kb) initiation zones (IZs), although narrower origins were also reported [7]. It was unclear if these variable results reflected the true genomic diversity of replication origins or different technical biases.

The advent of DNA microarrays and high-throughput sequencing has allowed much broader and systematic scrutiny of origins. Crucially, different pictures were obtained depending on the technique used to purify initiation intermediates. Small nascent strands (SNS) synthesized at origins were purified by size selection, followed by λ-exonuclease digestion of the contaminating broken DNA strands lacking a protecting 5'RNA primer (λ-SNS) [8,9], or by briefly labelling newly synthesized DNA with BrdU (5-bromo-2'-deoxyuridine) with BrdU (5-bromo-2'-deoxyuridine) or digoxigenin-dUTP, followed by size selection and immunoprecipitation [10-12]. Replication bubble-containing restriction fragments were purified

by trapping in gelling agarose and electrophoretic elimination of bubble-devoid fragments [13,14]. These independent approaches to purify initiation intermediates often gave poorly concordant origin locations. Furthermore, SNS tended to highlight site-specific origins whereas bubbles revealed broad initiation zones. Lastly, no information about fork progression and termination could be obtained by these approaches.

Replication timing (RT) and replication fork direction (RFD) profiling are orthogonal approaches to study DNA replication. They do not depend on isolating initiation intermediates but on analysis of the replication behaviours of all investigated loci. In Repli-seq, cells are pulse-labelled with BrdU for 30-120 min, sorted by FACS relying on total DNA content into 2-6 fractions of S phase. Next, BrdU-DNA is immunoprecipitated and hybridized to microarrays or sequenced, allowing to generate global RT profiles [15,16]. In a recent improvement called high-resolution Repli-seq, up to 16 fractions of S phase were used [17]. A second approach for measuring RT is based on assaying DNA copy number by sequencing sorted S and G1 cells, or even unsorted asynchronously proliferating cells [18]. Importantly, Repli-seq and copy number profiles are highly consistent with each other and extremely reproducible between laboratories [19]. They identify peaks and valleys of early- and late-replicating DNA, respectively, but unlike in yeast, their spatial and temporal resolution (~2 h and ~100 kb) is insufficient to precisely map origins in mammals. In high-resolution Repli-seq, however, the resolution was improved to 50 kb, allowing the identification of some isolated IZs [17].

Genome-wide RFD profiles were first obtained by nucleotide compositional skew (*S*) analysis, following the seminal observation of abrupt *S* sign change signs at bacterial origins and termini [20]. Analysis of mammalian genomes revealed ~1,000 abrupt *S* upshifts similar to those at bacterial origins, separated by megabase-sized segments of progressive *S* decrease, tracing N-shaped domains of *S* [21,22]. *S* accumulates during evolution due to mutational asymmetries between the leading and lagging strands [23]. *S* amplitude and sign, therefore, reflect the average fork direction in the germline. Comparison with RT profiles of somatic cells corroborated that replication progresses from the borders to the centres of N-domains, suggesting developmental and evolutionary conservation of these replication patterns [24-26]. However, many more origins than *S* upshifts were found in mammalian genomes; the missing origins must therefore be flexible enough or located within regions frequently rearranged to leave no evolutionary stable imprint on *S* profiles. Furthermore, the resolution was limited to

~20 kb and analysis of gene-rich regions was complicated by the added effect of transcription-associated mutational asymmetries [27]. These limitations called for a genome-wide, direct experimental determination of RFD at high resolution in mammalian genomes, which was first achieved by sequencing of purified Okazaki fragments [28].

Development and overview of OK-seq

At the replication fork, the leading strand is replicated continuously whereas the lagging strand is synthesized discontinuously, in the form of ~200 nt RNA-primed fragments (Okazaki fragments) that grow in the direction opposite to fork progression. Okazaki fragments are joined together one after another to build an elongating lagging strand. Okazaki fragments mapping to the Watson and Crick strands are generated by leftward- *(L)* and rightward- *(R)* moving forks, respectively. Therefore, strand-oriented sequencing of Okazaki fragments isolated from a cell population reveals the proportions of *R* and *L* forks at any locus, allowing quantitative analyses of replication fork initiation, progression, and termination. Isolation and sequencing of Okazaki fragments were first achieved in ligase- and checkpoint deficient mutants of *S. cerevisiae*, which allowed continued DNA synthesis despite the accumulation of unligated Okazaki fragments behind the forks [29,30]. We independently developed a procedure for isolating and sequencing Okazaki fragments from mammalian cells that did not require the introduction of such mutations. In this method, asynchronously growing cells are briefly pulsed with 5-ethynyl-deoxyuridine (EdU) to label newly-synthesized DNA, total DNA is denatured and fractionated by size, and the < 200 nt EdU-labelled single DNA strands are clicked with biotin, captured on streptavidin beads, and ligated to sequencing adapters. This procedure was dubbed OK-seq[28] (Fig. 2).

The replication fork directionality (RFD = $R - L$) profiles thus obtained were highly resolutive (~1 kb for human cells and ~50 bp for yeast) and informative. RFD at position x is mathematically linked to the mean RT (MRT) and to the speed of forks (v), such that $dMRT/dx = RFD/v$ [25,26]. In other words, steep MRT slopes correspond to unidirectionally replicating regions, flat MRT zones are replicated equally often in both directions, and intermediate MRT slopes are replicated by unequal proportions of *R* and *L* forks. Indeed, the human OK-seq RFD profiles were found to be extremely consistent with RFD profiles derived from skew and MRT data, but more resolutive. In yeast, RFD upshifts, where fork direction switches from *L* to *R*, span one kb or less, identifying site-specific origins (Fig. 1b and 3b) at locations highly consistent with previous origin mapping studies [30,31]; see below for details). However, a completely different RFD pattern is observed along the human

genome [28]. Most loci show a mixture of R and L forks, and changes in RFD are progressive rather than abrupt, spanning tens or hundreds of kb (Fig. 1b and 3a). These results imply an extensive cell-to-cell variability in replication patterns. An automated procedure based on a hidden Markov model (HMM)[28,32] was developed to objectively detect ascending, descending and flat RFD segments across the entire genome. Extended flat segments with extreme RFD values (close to ±1), which reveal unidirectionally replicating regions, only cover 5-10% of the genome. Segments of ascending RFD, where fork direction progressively shifts from *L* to *R* (IZs), typically span 10-100 kb. They reveal 4,000 – 10,000 IZs which support a low and homogeneous rate of initiations over their entire length. The amplitude of the shift reveals the global efficiency of each zone (i.e. the fraction of molecular copies which support an initiation event), which ranges from <10% to >90%. Abrupt upshifts such as those found at yeast origins are extremely rare. Descending RFD segments between consecutive IZs reveal extended (10 – 1000 kb) zones of replication termination (TZs), even broader than the IZs. Finally, extended segments of null RFD reveal randomly replicating regions, mostly in late-replicating heterochromatin [28].

Importantly, when OK-seq was adapted to purify EdU-labelled Okazaki fragments from *S. cerevisiae*, very similar profiles to those reported for ligase- and checkpoint-deficient *S. cerevisiae* mutants were obtained, consistent with the site-specific nature of yeast origins [31]. Therefore, the much broader RFD upshifts observed in mammalian genomes reflect the different biology of yeast and mammalian cells and not an inability of OK-seq to reveal abrupt RFD upshifts, characteristic of site-specific origins (Fig. 1 and 3).

Given the cell-to-cell variability and dispersed nature of replication initiation and termination events, particularly in mammalian cells, caution is required to interpret changes in RFD along the profiles. Strictly speaking, the ΔRFD between two genomic positions is equal to twice the *difference* between the number of initiation and termination events in the considered interval. For example, a segment across which the RFD continuously decreases from +1.0 to -1.0 may simply be invaded by outer forks that merge at variable positions, resulting in a single, delocalized termination event (Fig. 1c). However, a similar decreasing RFD segment may also arise if one internal, delocalized initiation event emits two diverging forks that meet at random positions with the two outer invading forks, resulting in two delocalized termination events. More generally, any scenario where *n* initiation and *n+1* termination events take place between the two outer invading forks can explain the decreasing RFD pattern, with a likeliness that increases with the size of the segment. Similarly, an ascending RFD segment

may in principle arise from *n* initiation events interspersed with *n*-1 termination events. However, ascending RFD segments are markedly smaller than descending ones, so the scenario with at most one initiation event and no termination event, as first demonstrated for the DHFR initiation zone [2], is by far the most likely explanation. Single-molecule replication analyses of the yeast genome [31] and two chicken chromosome fragile sites [33] recently confirmed that a minor fraction of initiation and termination events occur in negative and positive RFD slopes, respectively. In addition, recent high-throughput single-molecule optical replication mapping of early initiation events of human cells [34] also confirmed that a minor fraction of early initiation events occurs in negative RFD slopes as well as within late randomly replicating regions. Therefore, the positive or negative slope of an RFD segment reveals whether initiation or termination *predominates*, but a mixture of both, on different molecules or on the same molecule, cannot be excluded. Given that the number of ascending RFD segments in mammalian cells (4,000 – 10,000) is lower than the estimated number of initiation events per S phase (20,000 – 50,000) and that most IZs support at most one initiation event per cell cycle, the simplest model to reconcile these numbers is that many initiation and termination events occur within TZs and null RFD regions but in a too dispersed manner to leave an imprint on population RFD profiles. Such dispersed events can only be detected by single-molecule techniques [31,33,34]

Applications of OK-seq

OK-seq was used to obtain high-resolution, genome-wide RFD profiles of many types of cultured metazoan cells [28,33,35-38] and even of primary B cells stimulated to proliferate *in vitro* [39]. With the continuing development of novel origin mapping techniques, it should be noted that OK-seq IZs have been recently confirmed by EdU-seq HU [39], by high-resolution Repli-seq [17], and by optical replication mapping [34].

The HMM automated analysis of the RFD slope presented here allowed to map IZs and TZs and to measure their efficiencies [28]. Alternatively, IZs and TZs can be automatically detected in OK-seq profiles by Wavelet-Transform Analysis (WTA) [40]. IZs often but not always abut active genes, and are seldom if ever transcribed, consistent with reports that licensed origins are eliminated from transcribed genes [2,41-44]. Thus, there is a tight association of gene activity with replication initiation in the flanking intergene(s). IZs are often flanked on one or both sides by active genes. Due to the different strengths of the 5' and 3' IZs, however, active genes tend to be replicated in the same direction as transcription, as first observed by

nucleotide compositional skew analysis [22], but this is far from an absolute rule. In particular, the RFD tends to invert over long active genes such that their 3'end is often replicated in the direction opposite to transcription [45,46].

Whereas IZs bordering active genes fire in the early S phase, IZs remote from active genes fire later. Furthermore, the flat RFD segments found in late-replicating heterochromatin, support widespread, delocalized initiation in the absence of stimulating gene activity [28]. Perhaps surprisingly, however, RFD profiles are generally more variable between cell lines in the AT-rich, late-replicating, gene-poor isochores than in the GC-rich, early replicating, gene-rich isochores [35].

Besides replication program characterization of normal and cancer cells [28,35,36,39] and of cells subjected to replication stress [37], OK-seq has become very useful in a broad range of genomic studies. First, the inability to initiate replication within transcribed genes has been proposed as a mechanism for causing DNA breaks at common chromosomal fragile sites (CFSs) harbouring long genes due to delayed replication [46-48]. The identification of unidirectionally replicated regions by OK-seq, combined with MRT analysis, indeed allowed to predict CFSs genome-wide [46]. Second, the high probability of initiating replication between active genes in early-replicating domains was confirmed by EdU-seq HU [39]. This study found that DNA double-stranded breaks (DSBs) induced by S-phase entry in the presence of hydroxyurea are also confined between active genes, ruling out replication-transcription collisions as their cause. Early-replicating fragile sites (ERFs) instead represent zones where forks collapse after origin firing in hydroxyurea. Nucleosome-depleted, asymmetrical AT-rich motifs bordering initiation sites act as polar barriers to fork progression and trigger DSB accumulation. These results suggest a complex interplay between replication initiation and chromosome fragility [39]. Third, OK-seq data have been used to compare the density of MCM proteins, which mark potential replication origins, to the probability of initiation along the genome. The lack of initiation within transcribed genes was explained by a depletion of MCM proteins within gene bodies. However, ascending and descending RFD segments of similar RT and transcription status did not show different MCM densities, suggesting that additional factors to MCM density act to determine the probability of initiation along the genome [40]. Fourth, OK-seq data revealed that active genes tend to replicate codirectionally with transcription [28]. Later studies employing OK-seq data further revealed that head-on, but not codirectional, collisions between replication and transcription lead to the accumulation of potentially deleterious RNA-DNA hybrids (R-loops) [49], that replication stress markers

accumulate at transcription termination sites, where forks progress head-on to transcription, but not at transcription start sites, where forks progress codirectionally with transcription [45] and that numerous factors, such as topoisomerase 1 [45,50], the SAMHD1 ribonuclease [51] and the SWI/SNF chromatin remodelling complex [52] process R-loops and help resolve transcription-replication conflicts. Fifth, mapping RFD by OK-seq has contributed to revealing that leading and lagging strands are prone to different mutational rates across evolution and during cancer transformation, and have helped to deconvolve the strand-asymmetrical production of mismatches by leading and lagging-strand DNA polymerases from their strand-asymmetrical removal by mismatch repair [28,53-55]. OK-seq data have also contributed to reveal the strand-biased integration preferences of LINE-1 retrotransposons [56,57]. Sixth, combining OK-seq with strand-specific profiling of replicated chromatin demonstrated that inheritance of parental modified histones proceeds by distinct mechanisms at the leading and the lagging strands [36,38], and combining OK-seq with DNA remethylation analysis revealed that leading and lagging strands acquire DNA methylation with slightly different kinetics [58].

In sum, OK-seq is a quantitative method allowing to reveal the genome replication dynamics and the impact of DNA replication on genome and epigenome function and evolution.

Comparison with other methods

Other direct and indirect methods for measuring replication directionality have been developed by different groups. As discussed above, nucleotide compositional skew analysis [21,22] and spatial derivation of MRT profiles [25,26] gave RFD profiles highly consistent with, but less resolutive than OK-seq [28]. The enrichment of Okazaki fragments for direct sequencing was first achieved in *S. cerevisiae* through ligase and checkpoint inactivation [29]. While yeast RFD profiles obtained by this method and by OK-seq are extremely similar [31], the ligase-inactivation approach predominantly enriches for mature Okazaki fragments while the EdU-mediated purification enriches for growing Okazaki fragments, which is important to keep in mind when analysing Okazaki fragment processing and nucleosome phasing.

Recent indirect methods to map RFD are based on the fact that the leading (Pol ε) and lagging (Pols α and δ) strand replicative polymerases incorporate ribonucleotides into

genomic DNA at different rates. Ribonucleotide excision repair (RER) mutants are viable, and polymerase mutants that incorporate ribonucleotides at higher rates than wild-type have been obtained. Four methods (dubbed EmRiboSeq [59], Pu-Seq [60], HydEn-Seq [61]and Ribose-Seq [62] were reported to determine the genome-wide distribution of embedded ribonucleotides, and infer RFD, across the genome of RER and polymerase mutants in *S. cerevisiae* and *S. pombe*. They also identified regions in which ribonucleotide incorporation deviates from lagging/leading strand expectations, such as at replication origins, which were proposed to result from leading strand initiation by Pol δ followed by an exchange with Pol ε [60], and at termini, suggesting a reciprocal switch from Pol ε to Pol δ [63]. A recent preprint reported the extension of Pu-seq to human cells [64].

A new method for strand-specific sequencing of short nascent strands revealed that SNS are distributed with a sharp strand-specific asymmetry around the peak summits [65]. This finding is surprising as, during origin firing, SNS are expected to grow in both directions by leading and lagging strand synthesis from two forks.

Novel methods for mapping DNA breaks were reported to indirectly reveal RFD, suggesting that the frequency and/or kinetics of nick repair is distinct between the leading and lagging strands. The GLOE-seq method, which maps single-strand breaks in a strand-specific manner, also provided high-resolution RFD profiles in mammalian and yeast cells. GLOE-seq uses a reduced input cell number compared to OK-seq, yet it requires ligase inactivation [66]. A conceptually similar method that differs in library preparation strategy, TrAEL-seq, allows to map the 3' ends of double-strand breaks and provides RFD information [67].

Although the OK-seq approach is now well developed, up to date, there is no available bioinformatics protocol to fully explore the data. A recently published *Nature Protocol* paper [68], provided a simple approach to profile RFD around aggregate genomic features (such as transcription start sites), but no method to call IZ and TZ. Here, we provide a complete protocol for using an R-based toolkit, OKseqHMM (https://github.com/CL-CHEN-Lab/OK-Seq), to process and analyse OK-seq data, along the genomes of different species (human, mouse, and yeast). Following the current protocol, we can *(1)* visualize high-resolution RFD profiles (1 kb for human/mouse cells and 50 bp for yeast) and detect the IZs and TZs by using

a 4-state Hidden Markow Model (HMM), *(2)* calculate the origin efficiency metric (OEM) [30] and visualize RFD changes at different scales and *(3)* visualize the RFD and OEM profiles over genomic features of interest. This toolkit provides a useful resource for the broad scientific community working on DNA replication, genomic instability, and epigenetics.

Limitations

One limitation of OK-seq is that, as any cell population method, it averages cell-to-cell variability. As with other NGS-based origin mapping approaches, rare events cannot be directly seen. Although cell-to-cell variability remains visible since most loci show a mixture of *R* and *L* forks, dispersed initiation and termination events may go undetected even if they represent the majority of events. For example, long segments of null RFD can only be explained by random initiation and termination, but the density of these events cannot be measured. The change in RFD across a segment is equal to twice the difference between the number of initiation and termination events within the segment [69]. Therefore, a minority of termination (resp. initiation) events may occur within ascending (resp. descending) RFD segments, and only single-molecule methods may directly reveal them [31,33,70]. The OK-seq results thus led us to propose that replication of mammalian genomes combines predominant initiation within "master" IZs detected as ascending RFD segments, with more dispersed, less efficient initiation elsewhere.

OK-seq relies on the metabolic labelling with nucleotide analogs (EdU) and we anticipate that it may be used in any proliferating cells or even model organisms able to efficiently uptake EdU. OK-seq requires a significant amount of starting material since the half-life of Okazaki fragments is very short. Furthermore, the library preparation step may benefit from future improvements, for example inspired from single-stranded library preparation from ancient genomes [71], although optimization will be required.

Expertise needed to implement OK-seq

OK-seq requires strong skills in molecular and cell biology. The protocols are accessible to most molecular biology laboratories and rely on common laboratory equipment. Bioinformatic analysis with pre-built pipelines requires strong computational skills and experience with R.

Experimental design

Here we present some critical considerations and the key steps of the experimental and analytical workflows of OK-seq (Fig. 2).

Cell culture and starting cell number

Since we purify Okazaki fragments from unperturbed asynchronously growing cells, their amount is expected to be tiny, around hundreds of picograms per million asynchronous cells. This is why Okazaki fragment isolation requires a large number of input cells (3-10 × 10^8). This consists of large-scale cell cultures which needs to be carefully planned. Cell numbers may be optimized depending on genome size and fraction of cells in S phase. For example, a lymphoblastoid cell line of nearly normal karyotype with approximately 20% of cells in S phase (GM06990) required 8-10 × 10^8 cells per biological replicate, whereas hyperploid cancer cell lines with 30-35 % of cells in S phase, such as HeLa or K562, required 3 × 10^8 cells per replicate. Cell cultures should be split 1 or 2 days before the experiment, to ensure small colonies and uniform labelling.

EdU labelling and cell harvesting

In this step, newly-synthesized DNA strands are briefly labelled with ethynyl-containing nucleotide EdU [72]. The Okazaki fragments are transient and are immediately ligated to the elongating nascent lagging strands, with a half-life shorter than 10 seconds [73,74]. We set the EdU pulse for 2 minutes because it was easy to keep consistent between experiments at a comfortable working pace. Yet, in theory, the pulse could be shortened since thymidine analogs are almost instantly assimilated. In contrast, longer pulses will increase the proportion of nascent labelled DNA of higher molecular weight which could contaminate Okazaki fragment preparation. In any case, the duration of the pulse needs to be precisely controlled and stopped abruptly by adding ice-cold PBS. It is, therefore, preferable to treat a small number (2-3) of dishes at the same time. Option A of this section explains how to label and harvest adherent cells (HeLa) and option B explains how to treat the cells growing in suspension (EBV-immortalized lymphoblastoid GM06690). For labelling, we have also previously used a cytidine analog EdC [75], which in HeLa cells gave an identical result to EdU [28]. However, the use of EdC has limitations, as EdC assimilation efficiency varies in different cell types and depends on cytidine deaminase activity [76,77].

Nucleic acid extraction Nucleic acids are extracted with the proteinase K / phenol-chloroform method [78], which allows inexpensive milligram-scale preparation of pure high-molecular-weight genomic DNA. At this step, it is critical to avoid pipetting and vortexing to minimize DNA breakage and potential contamination of Okazaki fragment preparation with

fragments of elongating nascent strands. After ethanol precipitation, we typically leave the DNA pellet in TE buffer during 3-7 days at 4 °C to allow it to dissolve without pipetting. We omit RNAse A digestion and use intracellular RNAs as molecular cargo during subsequent purification steps.

Size-fractionation and recovery of small single-stranded fragments

To release Okazaki fragments, genomic DNA is heat denatured and size-fractionated on neutral linear 5-30 % sucrose gradients [79]. We use one 36 ml gradient to fractionate 500 µg of prepared genomic DNA (1-1.5 × 10^8 of starting cells), which represents 6-10 gradients per experiment. Sucrose gradients should be handled with caution. After overnight centrifugation, the fractions containing small fragments (< 250 nt) are concentrated and buffer-exchanged.

Biotinylation in click reaction

For isolation of EdU-labelled replicated DNA, EdU is coupled with biotin-TEG-azide in click reaction (Copper-Catalyzed Azide-Alkyne Cycloaddition (CuAAC) click chemistry) [80-82]. Afterwards, cellular RNAs, including the RNA portions of Okazaki fragments are hydrolyzed with alkali and 5' extremities of DNA fragments are phosphorylated with T4 PNK.

Sequencing adapter ligation and streptavidin capture of biotinylated fragments

In OK-seq, it is critical to prepare strand-oriented libraries from single-stranded DNA with a minimal technical bias, to achieve uniform coverage of reads over the genome. In library preparation, the double-stranded DNA ligation with T4 DNA ligase is used since it has the lowest sequence preference compared to single-stranded DNA ligation [83]. Two different double-stranded adapters with a single-stranded random hexanucleotide overhang are hybridized to the ends of the purified fragments. To reduce self-complementary interactions of 5' adapter (A1) and 3' adapter (A2), the standard Illumina sequence of 5' adapter was shortened by 5 bases [84]. To prevent self-ligation, adapter A2 contains 3'-terminal dideoxy-modifications (Table 1). After the ligation step, the library fragments containing nascent biotinylated molecules are captured with streptavidin-coated magnetic beads. We perform an additional step of hybridization and ligation of adapters on beads to increase the chance of successful recovery of Okazaki fragments into the library. Each step is followed by stringent high-salt washes to remove the non-specifically bound DNA molecules and unligated adapters.

Library amplification and sequencing

Libraries are amplified by PCR with indexing primers (Table 1). The template library

fragments remain attached to the beads during PCR and may be recovered, washed, and reused for an additional round of amplification. In our hands, this additional amplification step resulted in a much higher yield of the final amplified library with nearly identical library complexity, without a strong increase in PCR duplicates [28]. PCR products containing > 30 bp inserts are size-selected and eluted from agarose gels. Illumina sequencing is performed following standard protocols but replacing the sequencing primer of the first read by the shortened primer [84].

Data processing

The raw sequencing data (e.g. fastq files) need to be pre-processed and aligned to a reference genome using standard bioinformatics procedures. In our toolkit, the first function (OKseqHMM) automatically detects whether the input aligned sequencing data are single-end or paired-end reads, then splits reads into Watson and Crick strands and calculates the RFD values within adjacent windows (by default 1 kb) along the reference genome ; $RFD = \frac{C-W}{C+W}$, where *C* and *W* correspond to the number of reads mapped on the Crick and the Watson strands respectively. Next, HMM algorithm allows segmentation of RFD profile into upward, downward and flats segments to predict the location of initiation, termination and unidirectional fork movement zones respectively. The second function of the tool kit, OKseqOEM, uses the Watson and Crick strand aligned reads to compute the OEM at multiple scales defined by a user ; $OEM = \frac{W_L}{W_L+C_L} - \frac{W_R}{W_R+C_R}$ (where W_L and W_R are the number of reads in the left and right quadrants of the Watson strand while C_L and C_R refer to the read numbers in the left and right quadrants of the Crick strand. And finally, the function AveragePlot generates average metagene profiles and heatmaps to analyze the distribution of RFD and OEM around genomic features of interests.

MATERIALS

Biological materials

CRITICAL For the yeast *S. cerevisiae*, please refer to the supplementary protocol.

Human cell lines

- HeLa MRL2 (a kind gift from Dr Olivier Bensaude, IBENS)
- Immortalized lymphoblasts GM06990 (Coriell)

! CAUTION The cell lines used in your research should be checked regularly to ensure they are authentic and mycoplasma-free.

CRITICAL Use the appropriate medium and supplements for the cell type of interest.

Reagents

Cell culture reagents for HeLa cells

- DMEM (Gibco, Cat. No. 31966-021)
- Fetal bovine serum (FBS) (Sigma-Aldrich, Cat. No. F2442)
- Penicillin-Streptomycin (10,000 U/mL) (Gibco, Cat. No. 15140-122) **! CAUTION**. Irritant if contact with skin. Always wear gloves and a lab coat.
- Trypsin-EDTA (0.25 %) (Gibco, Cat. No. 25200-056)

Cell culture reagents for GM06990

- 1× PBS (Thermo Fisher, Cat. No. 14200083)
- Fetal bovine serum (FBS) Sigma-Aldrich, Cat. No. F2442
- Penicillin-Streptomycin (10,000 U/mL) (Gibco, Cat. No. 15140-122) **! CAUTION**. Irritant if contact with skin. Always wear gloves and a lab coat.
- RPMI1640 (Thermo Fisher, Cat. No. 61870127)

Common reagents

- 5 M Betaine (Sigma-Aldrich, Cat. No. B0300-5VL)
- 5-ethynyl-2'-deoxycytidine (EdC) (Jena Bioscience, Cat. No. CLK-N003-10) OPTIONAL (See experimental design).
- 5-Ethynyl-deoxy-uridine (5-EdU) (Jena Bioscience, Cat. No. CLK-N001-25)
- 50% PEG8000 (Jena Bioscience, Cat. No. CSS-256)
- Absolute ethanol (Sigma-Aldrich, Cat. No. 1117272500) **! CAUTION** Ethanol is flammable and irritative.
- Acetic acid (Sigma-Aldrich, Cat. No. 33209) **! CAUTION** Flammable, volatile, and irritative. Work under a chemical hood and wear gloves and a lab coat when handling.
- Agilent High Sensitivity DNA Kit (Agilent, Cat. No. 5067-4626)
- Ammonium acetate (VWR, Cat. No. 21200.297).) **! CAUTION** Work under a chemical hood while wearing a lab coat and disposable gloves.

- AMPure beads (Beckman, Cat. No. A63881)
- ATP (Thermo Fisher, Cat. No. R0441) **CRITICAL** Aliquot into 20-50 µL aliquots, store at -20 °C and avoid multiple freeze-thaw cycles.
- Biotin-TEG azide (Berry & associates, Cat. No. BT1085)
- Bromophenol blue (Sigma-Aldrich, Cat. No. 32712-5G) **! CAUTION** Work wearing a lab coat and disposable gloves.
- Chloroform (VWR, Cat. No. BDH83627.400) **! CAUTION** toxic and corrosive. Work under a chemical hood while wearing a lab coat and disposable gloves.
- Copper (II) sulphate ($CuSO_4$; Jena Bioscience, Cat. No. CLK-MI004-50) **! CAUTION**. It is irritative to the skin and eyes and is toxic if swallowed. Work wearing a lab coat and disposable gloves.
- Dimethyl sulfoxide (DMSO; Sigma-Aldrich, Cat. No. D2650) **! CAUTION** DMSO is harmful to the skin and is combustible. Work wearing a lab coat and disposable gloves.
- Distilled deionized water (ddH_2O) or UltraPure DNase/RNase-Free Distilled Water (Thermo Fisher, Cat. No. 10977035)
- dNTPs (Thermo Fisher, Cat. No. R0192) **CRITICAL** Prepare 5-10 µL aliquots, store at -20 °C and avoid freeze-thawing
- Dynabeads MyOne streptavidin T1 (Thermo Fisher, Cat. No. 65601)
- EB buffer (Qiagen, Cat. No. 19086)
- EDTA Ultrapure (0.5 M, pH 8.0; Life Technologies, Cat. No. 15575-038) **! CAUTION** Toxic if swallowed. Work wearing a lab coat and disposable gloves.
- Gel loading buffer II (2×, for UREA PAGE) (Thermo Fisher, Cat. No. AM8546G) **! CAUTION** It contains formamide and is toxic. Work under a chemical hood while wearing a lab coat and disposable gloves.
- Gel Loading Dye, Purple (6×, for PAGE and agarose gels) (NEB, Cat. No. B7024s)
- KAPA HiFi HotStart DNA Polymerase (Roche, Cat. No. 07958889001)
- Low Molecular Weight DNA Ladder (NEB, Cat. No. N3233S)
- Micro Bio-spin columns P30 (Biorad, Cat. No. 732-6250)
- MinElute Gel extraction Kit (Qiagen, Cat. No. 29604)
- MinElute PCR Purification Kit (Qiagen, Cat. No. 28004)
- 1 × PBS (Thermo Fisher, Cat. No. 14200083)
- 10 × PBS (Thermo Fisher, Cat. No. 70011044)

- Phenol chloroform isoamyl alcohol (25:25:1) (Thermo Fisher, Cat. No. 15593-049) **! CAUTION** Toxic and corrosive. Work under a chemical hood while wearing a lab coat and disposable gloves.
- Potassium Acetate (CH_3COOK; Calbiochem, Cat. No. 529553)
- Primers for sequencing adapters and library construction (Common supplier, Table 1).
- Proteinase K (Roche, Cat. No. 3115879001)
- Qubit dsDNA BR Assay Kit (2-1000 ng/µl) (Cat. No. Q32853)
- Qubit ssDNA HS Assay Kit (0.05-100 ng/µl) (Cat. No. Q10212)
- Small Fragments Agarose (Eurogentec, Cat. No. EP-0020-10)
- Sodium acetate (Merck, Cat. No. 1.06268.0250).
- Sodium ascorbate (Jena Bioscience, Cat. No. CLK-MI005-50)
- Sodium chloride (NaCl) (Sigma-Aldrich, Cat. No. S7653)
- Sodium dodecyl sulfate (SDS) solution 20% (wt/vol) (Sigma-Aldrich, Cat. No. 05030-500ML-F) **! CAUTION** SDS is corrosive to the skin and a respiratory irritant. Work wearing a lab coat and disposable gloves. Thoroughly wash with water any skin or eyes exposed to this chemical.
- Sodium hydroxide NaOH (Sigma-Aldrich, Cat. No. 1.06469.1000) **! CAUTION** NaOH is corrosive. Wear gloves and a lab coat when handling.
- Sucrose (Sigma-Aldrich, Cat. No. 1.07687.5000)
- SYBR™ Gold Nucleic Acid Gel Stain (10,000 × Concentrate in DMSO) (Thermo Fisher, Cat. No. S11494) **! CAUTION** It is a potential cancer hazard. Work wearing a lab coat and disposable gloves.
- SYBR™ Green I Nucleic Acid Gel Stain 10,000 × (Thermo Fischer, Cat. No. S7585) **! CAUTION** It is a potential cancer hazard. Work wearing a lab coat and disposable gloves
- T4 DNA ligase (Thermo Fisher, Cat. No. EL0014) **CRITICAL** Aliquot the ligase buffer into 20-50 µL aliquots. Store at -20 °C and avoid freeze-thaw cycles > 3.
- T4 polynucleotide kinase, T4 PNK (Thermo Fisher, Cat. No. EK0031)
- TAE buffer (Thermo Fisher, Cat. No. 15558026)
- Taq DNA polymerase (NEB, Cat. No. M0273)
- TBE buffer (Thermo Fisher, Cat. No. B52) **! CAUTION** Harmful if swallowed or inhaled. Work wearing a lab coat and disposable gloves.

- TBE Gels (10%; Thermo Fisher, Cat. No. EC62752BOX) **! CAUTION** The polyacrylamide gel is a potential cancer hazard. Work wearing a lab coat and disposable gloves.
- TBE-UREA Gels (10%; Thermo Fisher, Cat. No. EC68752BOX) **! CAUTION** The polyacrylamide gel is a potential cancer hazard. Work wearing a lab coat and disposable gloves.
- Tris-HCl buffer (1 M, pH 7.5; Thermo Fisher Scientific, Cat. No. 15567027)
- Tris-HCl buffer (1 M, pH 8.0; Thermo Fisher Scientific, Cat. No. 15568025)
- Tris (3-hydroxypropyl-triazolyl methyl) amine (THPTA) (Sigma-Aldrich, Cat. No. 762342) **! CAUTION** Skin and eye irritant. Work wearing a lab coat and disposable gloves.
- Triton X-100 (Molecular-biology grade; Sigma-Aldrich, Cat. No. T8787-100ml) **! CAUTION** Skin and eye irritant. Work wearing a lab coat and disposable gloves.
- Tween 20 (Sigma-Aldrich, Cat. No. P1379)
- HiSeq 3000/4000 SBS Kit (50 cycles) (Illumina, Cat. No. FC-410-1001)

Equipment

- 0.2-mL PCR tube (Eppendorf, Cat. No. 0030124332)
- 1.5-mL Eppendorf tube (Eppendorf, Cat. No. 33290)
- 2100 Bioanalyzer Instrument (Agilent, Cat. No. G2939BA)
- Allegra® 64R High-Speed Centrifuge (Beckman, 367588) with fixed angle rotor JLA-10.500 (Beckman, Cat. No. 369681)
- Amicon Ultra-15 Centrifugal filter Unit (Millipore, Cat. No. UFC901024)
- ART™ Wide Bore Filtered Pipette Tips, 1-mL (Thermo Fisher, Cat. No. 2079G)
- Beckman Coulter 25 × 89mm Ultra clean tube (Beckman, Cat. No. 344058)
- Benchtop centrifuge, refrigerated fixed angle rotor (Eppendorf, model no. 5424R)
- Benchtop centrifuge, swing bucket (Eppendorf, model no. 5910)
- Blades (Sigma-Aldrich, Cat. No. Z290947)
- Cell culture incubator (37 °C, 5% CO_2)
- Cell scrapers (Duscher, Cat. No. 010155)
- Counting chambers: KOVA Glasstic Slide 10 With Counting Grids (KOVA International, Cat. No. 87144) (or Hemacytometer or cell counter)
- Falcon® Tissue Culture Dishes 150 mm (VWR, Cat. No. # 25383-103)

- Falcon® Petri Flasks 175 cm2 (Corning, cat. # 353112)
- Falcon conical tubes 50ml Cellstar® (Greiner bio-one, Cat. No. # 227-261)
- Falcon conical tubes 15ml Cellstar® (Greiner bio-one, Cat. No. # 188-271)
- 500-mL centrifuge bottles (Beckman, Cat. No. 361691)
- DiaMag Rotator (Diagenode, Cat. No. B05000001)
- DNA LoBind® Tubes, 1.5-mL (Eppendorf, Cat. No. 022431021)
- DynaMag-2 Magnet (Thermo Fisher, Cat. No. 12321D)
- Eppendorf ThermoMixer® C (Eppendorf, Cat. No. EP5382000023).
- Evaporator (Eppendorf, Model No. 5301)
- Glass Pasteur pipettes (VWR, Cat. No. 14673-043; clean and autoclaved)
- Gradient maker (Hoefer, Cat. No. SG50) or Gradient Master (Biocomp, Cat. No. 108)
- Electrophoresis system, vertical (Hoefer, Model No. SE260-10A-1.5)
- Electrophoresis system, horizontal (Bio-rad, Model No. Sub -Cell Model 96)
- HiSeq 3000 System (Illumina, Cat. No. SY-401-3001) or equivalent.
- Integra Biosciences Pipetboy Accu 2 Pipette Controller (Fisher Scientific, Cat. No. 10798252)
- Laminar flow hood (ESCO, Model No. LVG-4AG-F8)
- Safe Imager 2.0 Blue-Light Transilluminator (ThermoFisher, Cat. No. G6600)
- Phase lock gel light 50-mL (5 Prime, Cat. No. 713-2539) or MaXtract High-Density 50-mL (Qiagen, Cat. No. 129073) or equivalent.
- 50-mL plastic pipettes (Corning, Cat. No. 07-200-17)
- 25-mL plastic pipettes (Corning, Cat. No. 07-200-15)
- 10-mL plastic pipettes (Corning, Cat. No. 07-200-12)
- ProFlex PCR System (Thermo Fisher, Cat. No. 4484073)
- Qubit 4 fluorometer (Thermo Fisher, Cat. No. Q33238)
- Qubit assay tubes (Thermo Fisher, Cat. No. Q32856)
- Sorenson low binding aerosol barrier tips, MicroGuard G, maximum volume 10 µL (Sigma-Aldrich, Cat. No. Z719374)
- Sorenson low binding aerosol barrier tips, MultiGuard, maximum volume 200 µL (Sigma-Aldrich, Cat. No. Z719447)
- Sorenson low binding aerosol barrier tips, MultiGuard, maximum volume 20 µL (Sigma-Aldrich, Cat. No. Z719412)

- Sorenson low binding aerosol barrier tips, MultiGuard, maximum volume 100 µL (Sigma-Aldrich, Cat. No. Z719463)
- SYBR Green I Nucleic Acid Gel Stain - Thermo Fisher (S7593)
- Optima XE-100-IVD Ultracentrifuge (Beckman, Part No. A99836) with swinging rotor SW28 (Beckman, Part No. 369650) or SW32 (Beckman, Part No. 342207)
- Vortex-Genie 2 (Scientific Industries, Cat. No. SI-A256)
- 250-mL glass beaker (clean and autoclaved, Fisher Scientific, Cat. No. FB101250)
- 600-mL glass beaker (clean and autoclaved, Fisher Scientific, Cat. No. FB101600)

Software

- deepTools (https://deeptools.readthedocs.io/en/develop/index.html) [85]
- IGV (https://software.broadinstitute.org/software/igv/) [86]
- OKseqHMM (https://github.com/CL-CHEN-Lab/OK-Seq)
- R (https://www.r-project.org/) [87]
    - R package "HMM" [88]
    - R package "Rsamtools" [89]
    - R package "GenomicAlignments" [90]
- Rstudio [91]
- wigToBigWig (http://hgdownload.soe.ucsc.edu/admin/exe/)

Reagent setup

CRITICAL Common stock solutions are prepared following standard molecular biology recipes [78] and ***http://cshprotocols.cshlp.org/site/recipes/nav_s.dtl***

Cell culture

- **DMEM-serum media for HeLa cells:**

  Mix 500 mL of DMEM medium with 50 mL of FBS, 5 mL of 100 × Penicillin-Streptomycin. Store at 4 °C for up to 2 weeks. Prewarm to 37° C before use.

- **RPMI 1640-serum media for GM06990 cells:**

Mix 500 mL of RPMI medium with 75 mL of FBS, 5 mL of 100 × Penicillin-Streptomycin, and 3.5 µL of ß-mercaptoethanol. Filter to sterilize and store at 4 °C for up to 2 weeks. Prewarm to 37° C before use.

Common reagents

- **100 mM Biotin-TEG azide**

  Dissolve 25 mg in 0.562 mL of DMSO. Store at 4 °C for up to 1 year.

- **100 mM $CuSO_4$**

  Dissolve 100 mg in 6.27 mL of ddH_2O. Aliquot and store at 4 °C for up to 1 year

- **1 M sodium ascorbate**

  Dissolve 200 mg in 1.01 mL of ddH_2O. Aliquot and store at -20 °C for up to 1 year**.**

CRITICAL Discard and prepare fresh if the solution has turned yellow.

- **20 mM EdU**

  Dissolve 25 mg in 4.956 mL of DMSO. Aliquot and store at -20 °C for up to 1 year.

- **2 × BWT**

  Prepare as outlined below. Can be stored at room temperature (RT; 22 °C) for up to 6 months.

Reagent	Final	Stock	Volume (mL) for 50 mL
Tris HCl pH 7.5	10 mM	1 M	0.5
EDTA pH 8.0	1 mM	0.5 M	0.1
NaCl	2 M	5 M	20
Tween 20	0.1 % (vol/vol)	10 % (vol/vol)	0.5
ddH_2O			Up to 50 mL

- **1 X BWT**

  Mix 25 mL of 2 × BWT with 25 mL of ddH_2O. Can be stored at RT for up to 6 months.

- **500 mM THPTA**

  Dissolve 100 mg in 460.3 µL of ddH_2O. Aliquot and store at 4 °C for up to 1 year.

- **80 % (vol/vol) ethanol**

  Mix 8 mL of absolute ethanol with 2 mL of ddH_2O. **CRITICAL** Prepare freshly each time.

- **AMPure XP beads**

  Divide bead solution into 2 mL aliquots and store at 4 °C (can be done in advance).

CRITICAL It is critical to equilibrate the AMPure XP beads at RT (≥ 22 °C) for at least 30 min before use for an optimal size selection.

- **DNA lysis buffer**

  Prepare as outlined below. Autoclave and store at RT for up to 1 year.

Reagent	Final	Stock	Volume (mL) for 500 mL
Tris-HCl pH 8.0	10 mM	1 M	5 mL
EDTA pH 8.0	25 mM	0.5 M	25 mL
NaCl	100 mM	5 M	10 mL
ddH_2O			Up to 500 mL

- **5% TEN-sucrose buffer**

  Prepare as outlined below. Autoclave and store at RT for up to 6 months.

Reagent	Final	Stock	Volume (mL) for 1 L
Tris-HCl pH 8.0	10 mM	1 M	10 mL
EDTA pH 8.0	1 mM	0.5 M	2 mL
NaCl	100 mM	5 M	20 mL
Sucrose	5 % (wt/vol)	50 % (wt/vol)	100 mL
ddH_2O			Up to 1000 mL

- **30% TEN-sucrose buffer**

  Prepare as outlined below. Add several crystals of bromophenol blue. Autoclave and store at RT for up to 6 months.

CRITICAL The bromophenol blue is optional but is very useful for gradient visualization.

Reagent	Final	Stock	Volume (mL) for 1 L
Tris-HCl pH 8.0	10 mM	1 M	10 mL
EDTA pH 8.0	1 mM	0.5 M	2 mL
NaCl	100 mM	5 M	20 mL
Sucrose	30 % (wt/vol)	50 % (wt/vol)	600 mL
ddH_2O			Up to 1000 mL

- **1 × TE-Tween**

Prepare as outlined below. Store at RT for up to 1 year.

Reagent	Final	Stock	Volume (mL) for 50 mL
Tris-HCl pH 8.5	10 mM	1 M	0.5
Tween 20	0.05 % (vol/vol)	10 % (vol/vol)	0.25
EDTA	1 mM	0.5 M	0.1
ddH_2O			Up to 50 mL

- **Oligonucleotides (Table 1)**

Order the primers listed in table 1 from a standard lab supplier. Adapters should contain the indicated modifications and be ordered in HPLC-grade, PCR primers can be ordered in a standard purification grade. Dissolve the oligonucleotides in EB buffer to the final concertation of 100 µM. Prepare working solutions of PCR primers by further diluting with nuclease-free H_2O to 10 µM. Store at -20 °C for up to 2 years.

CRITICAL The index sequences in the TruSeq Primers (lower-cased) can be substituted with any other index sequences. Double-indexing can be included in the primer sequences if desired. Primer R1 has an identical sequence to $A1_{top}$ but for naming simplicity is listed under a separate name.

PROCEDURE

Cell culture, EdU labelling, and cell harvesting ● TIMING 2-7 days of cell culture, 2 hours of labelling and harvesting.

1. Follow option (A) for adherent cells (Hela) and option (B) for suspension cells GM06990 (B).

CRITICAL For the yeast *S. cerevisiae*, please refer to the supplementary protocol.

(A) Cell culture, EdU labelling, and harvesting of adherent cells (HeLa)

(i) Culture adherent HeLa cells in 15-cm dishes with 20 mL of DMEM supplemented with 10% (vol/vol) FBS.

(ii) Seed 4×10^6 cells in 150-mm dishes with 20 mL of medium and grow them for approximately 48 h at 37 °C, 5 % CO_2 to reach 70-80% confluency. Prepare enough plates to harvest at least 500 million cells per one replicate (approximately 20 of 150-mm plates for HeLa cells).

CRITICAL It is important to respect the optimal cell culturing conditions and the density to maintain exponential cell growth.

CRITICAL The number of plates will depend on the cell size and seeding density. Additionally, the cell number may need to be optimized depending on the fraction of cells undergoing S phase in population and the cell ploidy. See Experimental design for the details.

(iii) Transfer 10 mL of the medium from the plate to a 50-mL tube and add 20 µL of 20 mM EdU stock solution. Mix by inverting the tube and pouring the EdU-containing medium back to the plate. The final EdU concentration is 20 µM. Incubate plates at 37 °C for exactly 2 minutes.

CRITICAL To keep the labelling time consistent between the plates, the EdU-containing medium has to be added and removed exactly in the same order and at a fixed time interval (e.g. 30 sec to 1 min) between plates. For convenience, we do not recommend handling more than 2-3 plates at the same time.

(iv) Aspirate the medium and immediately add 10 mL of ice-cold 1 × PBS to stop the EdU incorporation. Store the plates at 4 °C until all plates are processed.

(v) Collect the adherent cells by scraping with a clean cell scraper and transfer the cell suspension to 50-mL conical centrifuge tubes chilled on ice. To collect the remaining cells, rinse each plate with 10 mL of ice-cold 1 × PBS, and transfer to the same 50-mL conical tubes. Centrifuge for 10 min at 4 °C, 300 × g. Discard the supernatant.

PAUSE POINT. At this step cell pellets can be snap-frozen in liquid nitrogen and stored at -80 °C for up to one year.

(B) Cell culture, EdU-labelling, and harvesting of suspension cells (B-lymphoblasts GM06990)

(i) Culture cells in 175-cm^2 flasks with 50 mL of RPMI1640 media supplemented with 15% FBS, Penicillin/Streptomycin, and beta-mercaptoethanol at 0.8-1 million cells per mL.

(ii) Seed 2-2.5 × 10^7 cells in a 175-cm^2 flask vertically with 100 mL of medium for approximately 48 h at 37 °C, 5 % CO_2 to reach 0.8-1 million per mL. Prepare enough flasks to harvest at least 800 million cells per one replicate (8-10 flasks of

175 cm^2 for GM06990 cells).

CRITICAL It is important to respect the optimal cell culturing conditions and the density to maintain exponential cell growth. Lymphoblastoid cells make clumpy colonies at the bottom of the flasks. To maintain healthy cultures, resuspend the clumpy colonies to achieve single-cell suspension between the passages.

(iii) Carefully remove 80 mL of medium from the top of the flask using a pipette without disturbing cell clumps formed at the bottom of the flask. Save 20 mL of the medium in a 50-mL conical tube.

CRITICAL This step allows reducing the volume of the labelling medium. Lymphoblastoid cells form clumpy colonies on the bottom of the flask and the excess of the medium can be removed by aspirating from the top. For cell types growing in spinning flasks, cells can be centrifuged before the labelling and resuspended in a smaller volume of prewarmed medium.

(iv) Add 40 µL of 20 mM EdU stock solution to the 20 mL of the medium. Mix by inverting the tube and pouring the EdU-containing medium back to the flask containing 20 mL of cell suspension. The final EdU concentration is 20 µM. Incubate flasks at 37 °C for exactly 2 minutes.

(v) Cool the flasks by immerging and agitating them in an ice-cold water bath. Add 40 ml ice-cold 1 × PBS and 250 µL of 0.5 M EDTA, mix well. Store the flasks in the ice-cold water bath until all flasks are processed.

CRITICAL Respect the exact labelling time and immediately cool the flasks to quickly terminate the labelling.

(vi) Transfer cells to 50-mL Falcon tubes and centrifuge for 10 min at 4 °C, 300 × g. Discard the supernatant.

(vii) Resuspend all the pellets with 20 ml ice-cold 1 × PBS in one 50-ml Falcon tube. Centrifuge at 300 × g for 10 min at 4 °C. Discard supernatant.

PAUSE POINT. At this step cell pellets can be snap-frozen in liquid nitrogen and stored at -80 °C for up to one year.

Extraction of the genomic DNA • **TIMING** 2 hours with overnight incubation

CRITICAL For *S. cerevisiae* cells, follow the **Extraction of genomic DNA** section in the supplementary protocol.

2. Thaw the cell pellets on ice.
3. Resuspend cells in Lysis buffer to 1 million cells per mL. Distribute 10 mL aliquots of cell suspension to 50-mL tubes. Place the tubes on a rack at room temperature.

CRITICAL Gently resuspend to minimize cell rupture and DNA shearing. Achieve the single-cell suspension to ensure homogeneous lysis and optimal DNA extraction.

4. Add 250 µL of 20 % SDS to the cell suspension. The final SDS concentration is 0.5 % (wt/vol). Tightly close the cap and mix by gently inverting the tubes 5-10 times.

CRITICAL Keep the tubes at room temperature during SDS addition. Invert the tubes gently to minimize DNA breaks.

5. Add 50 µL of proteinase K 20 mg /mL (wt/vol) to the lysate. The final concentration of proteinase K is 0.1 mg / mL (wt/vol). Close the cap, and mix by gently inverting the tube.

CRITICAL At this stage the lysates will appear very viscous.

6. Incubate the tubes at 42 °C for 4 h or overnight (16 h).

CRITICAL After complete cell lysis the solution should appear homogeneous and transparent.

? TROUBLESHOOTING

7. In a chemical hood, add to each tube 10 mL of phenol-chloroform isoamyl alcohol mix solution pre-equilibrated at RT. Tightly close the cap and mix gently by inverting the tube until obtaining an entirely homogeneous mixture.

CRITICAL Bring the phenol-chloroform isoamyl alcohol solution to RT in advance.

CRITICAL Gently invert the tubes to allow the liquid to move between the cap and the bottom. This step may require up to 10 minutes.

! CAUTION Perform the DNA extraction inside a chemical hood and wear a lab coat and dispensable gloves.

8. Spin a 50-mL MaXtract High-Density tube at 1500 × g at RT for 2 min, and pour the mixture from Step 7 into the tube.
9. Centrifuge for 4 min at 1500 × g at RT with a swing rotor. This will separate the aqueous solution containing DNA while the organic phase will remain locked under the solid MaXtract gel phase.

CRITICAL Use of MaXtract High-Density tubes (or equivalent) is strongly recommended for achieving high-quality DNA preparation.

10. In a chemical hood, add to each tube 10 mL of phenol-chloroform-isoamyl alcohol mix. Tightly close the cap and mix gently by inverting the tube until full homogenization.

CRITICAL It is important that the organic fraction from step 9 remain locked under the MaXtract gel phase during this step.

11. Centrifuge for 4 min at 1500 × g at RT. This will separate the aqueous solution containing DNA while the organic phenol phase will remain locked under the solid MaXtract gel phase.

CRITICAL If the aqueous phase after this step is not clear, perform additional phenol-chloroform extraction by repeating steps 7-9.

12. In the chemical hood, add to each tube 10 mL of chloroform. Tightly close the cap and mix gently by inverting the tube until full homogenization. Centrifuge for 4 min at 1500 × g at RT.
13. Transfer the upper aqueous phase containing genomic DNA from all tubes by pouring into a clean 200-mL glass beaker.

CRITICAL Discard the organic fraction and the tubes to the appropriate chemical waste.

14. Add 2 mL of 7.5 M ammonium acetate per each 10 mL of lysate and mix gently with a Pasteur pipette.
15. Add 25 mL of absolute ethanol per each 10 mL of lysate, swirl gently with the same glass Pasteur pipette until the DNA precipitates.
16. Twine the precipitated DNA fibres with the Pasteur pipette and carefully transfer all the DNA precipitate into a clean 200-mL glass beaker containing 100 mL of 75 % (vol/vol) of ethanol. Leave the DNA precipitate immersed for 3~5 min. Repeat this step twice.

CRITICAL STEP It may be convenient to recover the DNA precipitate using two Pasteur pipettes as chopsticks.

17. Place the DNA precipitate with the Pasteur pipettes inside a new 15-mL falcon.
18. Remove any residual ethanol with a 1-mL tip.

CRITICAL STEP It is important to remove as much ethanol as possible.

19. Transfer the DNA precipitate to a new 15-mL tube, add 6 mL of TE.

CRITICAL STEP Ensure the entire DNA precipitate is immersed in TE buffer. Do not pipette.

20. Leave the Falcon open for 30 min at 37 °C in a dry thermostat to allow the evaporation of residual ethanol.
21. Carefully remove the Pasteur pipette and close the cap.

PAUSING POINT Store the DNA solution at 4 °C for at least 3-7 days to allow the complete dissolution of the DNA precipitate. Can be stored for up to 1 month at 4 °C.

Size-fractionation of denatured genomic DNA on neutral sucrose gradients. • TIMING 3.5 h of handling and 17 h of centrifugation.

CRITICAL STEP Because the centrifugation lasts 17 h it is convenient to start this step in the late afternoon.

22. Incubate the DNA solution from step 21 at 37 °C for 1 h to diminish the viscosity.
23. Measure the DNA concentration with Qubit ds DNA BR Kit according to the manufacturer's protocol. Typically, a yield of 2~3 mg of total DNA is expected.

? TROUBLESHOOTING

24. Split the volume into 6 equal aliquots of approximately 1-1.2 mL into 1.5-mL tubes using a 1-mL wide-bore tip.

CRITICAL STEP If the yield of total DNA is higher than 3 mg it is recommended to scale up the number of aliquots and gradient centrifugations accordingly.

CRITICAL STEP The DNA solution is viscous and hard to pipette at this stage. Pipette slowly with a 1-mL wide-bore to minimize DNA shearing.

25. Prepare 6 linear sucrose gradients in Beckman Coulter Ultra clear tubes 25 × 89 mm by mixing 18 mL of 5 % TEN-sucrose and 18 mL of 30 % TEN-sucrose using a gradient maker and following the gradient manufacturer's instructions.
26. Place each tube containing the gradients in a centrifuge tube adapter (Beckman Ultra-high-speed centrifuge, Rotor SW28) and hold it carefully. Proceed immediately to the next step.

CRITICAL STEP Due to the bromophenol blue in 30 % TEN-sucrose, a gradient of blue shade from the bottom to the top should be visible in the tube. If the blue gradient is not visible, discard the tube. Both Hoefer SG50 Gradient maker and Gradient Master (Rotor: SW28; Program: Long_Sucr_05-30%_wv_1St) result in similar and acceptable size-fractionation. We prefer Gradient Master as up to 6 highly uniform gradients can be simultaneously prepared within 15 min. Handle the gradients with caution.

27. Heat DNA aliquots from step 24 for 5 min at 94 °C to denature double-stranded DNA and chill immediately on an ice-cold water bath for 10 min.
28. Very carefully layer one aliquot of DNA from step 27 on the surface of one gradient from step 26 using a wide-bore tip. Load all gradients.
29. Adjust the weight of the tubes (with adapter) at symmetric positions on the rotor (1 and 4; 2 and 5; 3 and 6). Balance the weight by careful dropwise pipetting along the inner wall of the tube of the necessary amount of 5 % TEN-sucrose to achieve the exact (≤0.1 g) weight balance.

CRITICAL STEP Any minor imbalance may lead to the tube or the rotor breakage.

CRITICAL STEP Proceed immediately to the next step to avoid diffusion of the gradient.

30. Carefully close the caps, attach the adapters to the SW28 rotor and insert the rotor inside of the Beckman Ultracentrifuge. Spin under the vacuum for 17 h at 26,000 rpm at 20 °C, with acceleration and deceleration speed set on "High".

CRITICAL STEP Keep an eye on the centrifuge for about 15 min after the program starts to ensure that the desired centrifuge speed has been achieved.

PAUSE POINT Centrifugation lasts 17 hours.

31. The next day once the centrifugation is finished, switch off the vacuum and open the lid.
32. Carefully transfer the adapter with the tubes to the rack. Open the adapter lids with care.

CRITICAL Before collecting fractions, check the tube integrity. If the tube was broken during centrifugation the gradient should be discarded.

33. Number 18 of 15-mL Falcon tubes from 1 to 18.
34. Start collecting 1-mL fractions with a 1-mL wide-bore tip from the top of each gradient by slowly aspirating from the surface of the gradient. Combine fractions of the same order from all six gradients into a single 15-mL tube.

CRITICAL To collect the fractions, place a wide-bore tip vertically against the gradient surface and pipette slowly. Only pipette up from the surface of the gradient and never pipette down.

CRITICAL Usually the first 8 top 1-mL fractions contain DNA fragments of the desired size (≤ 250 nt), but we suggest collecting more fractions to check the size distribution and linearity of the gradient fractionation (Box 1).

CRITICAL Observe the colour of the fractions. Because of bromophenol blue in the dense sucrose solution, the top fractions should be lighter and the bottom fractions should appear progressively more coloured.

CRITICAL If wide-bore tips are not available, cut the 1-mL tips with clean scissors. Make sure the cut end is smooth and flat.

PAUSE POINT. The fractions can be stored at 4 °C for 1-3 days or frozen at -20 °C for up to 6 months.

35. Pool the fractions from step 34 containing fragments smaller than 200-250 nt (typically the first 1 to 8 fractions).
36. Concentrate the pooled fractions (48-80 ml) on a Millipore Amicon Ultra Centrifugal Filter, 15-mL, 10K.
37. Add 15 ml of fractions to a centrifuging filter and centrifuge at 4000 × g at room temperature for 10-15 min.
38. Discard the flow-through and load the next 15 ml of the sample to the filter. Repeat centrifugations until the entire volume of fractions is concentrated to approximately 300 µl.
39. Buffer-exchange by adding 5 ml of ultrapure water and centrifuge at 4000 × g for 10 minutes. Discard the flowthrough. Repeat 2 more times.
40. Transfer the concentrated solution from the filter (approximately 300 µl) to a new 1.5-mL tube. Measure the volume carefully with the pipette tip and note it on the tube.

PAUSE POINT. The concentrated fractions can be stored at -20 °C for 2 weeks.

? TROUBLESHOOTING

Click biotinylation • TIMING 2 h

41. Set up the click reaction by adding the following reagents sequentially to the tubes containing purified gradient fractions from step 40.

CRITICAL STEP If the volume of concentrated fractions from step 40 is > 375µl, scale up the volumes of all reagents.

Reagent	Volume (µL)	Final
DNA	≤ 375 µL	
10× Click-it buffer (or 10 × PBS pH 7.4)	50 µL	1 ×
100 mM Biotin -TEG- azide	5 µL	1 mM

500 mM THPTA	10 µL	10 mM
100 mM $CuSO_4$	10 µL	2 mM
100 mM sodium ascorbate	50 µL	10 mM
ddH_2O	Up to 500 µL	

42. Mix by pipetting and incubate for 45 min at RT.

CRITICAL The THPTA and $CuSO_4$ should be premixed and added in a single pipetting step.

CRITICAL Use freshly-prepared sodium ascorbate and respect the optimal reaction temperature. If the room temperature is lower than 22 °C, place the reactions in a thermoblock at 25 °C without mixing.

43. Spin briefly and split the 500 µL DNA solution into 2 equal aliqoutes of 250 µL in 2 1.5-mL Eppendorf tubes. Add 750 µL of absolute ethanol to precipitate DNA, close the caps and mix by inverting.
44. Chill the tubes at -80 °C for 15 min.
45. Spin for 30 min at ≥ 15,000 g) at 4 °C. Decant the supernatant.

CRITICAL The pellet can appear coloured in blue or brownish, probably due to the copper residue which does not interfere with the experiment.

46. Add 500 µL of 75% ethanol to the pellet, spin for 5 min at full speed at 4 °C. Decant the supernatant.
47. Quick spin and carefully remove the residues with a 200-µL tip without disturbing the pellet. Keep the tube open and air dry briefly (usually 2-5 min).
48. Dissolve each pellet in 45 µl of nuclease-free water and combine into a single 1.5-mL tube.

RNA Hydrolysis • TIMING 20 min

49. Add 10 µL 2.5 M NaOH into the 90 µL DNA from step 47 to a final concentration of 250 mM, mix by pipetting, quick spin, and incubate for 30 min at 37 °C
50. Quick spin and add 10 µL 2.5 M Acetic acid to neutralize the pH and mix by pipetting.
51. Purify the DNA with 2 × Biorad Micro Biospin P-30 columns according to the manufacturer's instructions.

52. Combine the purified flow throughs from two columns in one 1.5-mL tube.
53. Carefully measure the volume of the solution using a 200-µL pipette tip, and place the tube on ice.

CRITICAL Usually 120-150 µL DNA solution is recovered and the volume may slightly differ.

DNA phosphorylation and precipitation • TIMING 1.5 h

54. Set up the phosphorylation reaction by adding the following reagents sequentially to the tubes containing purified DNA from step 53. Mix by pipetting, quick spin, and incubate at 37 °C for 20 min.

Reagent	Volume (µL)	Final
DNA	≤ 117 µL	
10 × T4 PNK buffer A	15 µL	1 ×
10 mM ATP	15 µL	1 mM
T4 PNK (10 U/µL)	3 µL	0.2 U/µL
ddH2O	Up to 150 µL	

CRITICAL STEP If the volume of the DNA from step 53 is > 117 µl, scale up the volumes of all reagents accordingly.

CRITICAL STEP It is important to use a fresh aliquot of ATP and avoid freezing-thawing cycles.

55. Incubate the tube for 10 min at 75 °C to inactivate the T4 PNK enzyme.
56. Quickly spin the tubes and chill on ice.
57. To precipitate DNA, add 15 µL of 3 M sodium acetate (pH5.2) and 415 µL of -20 °C chilled absolute ethanol, mix by inverting. Incubate for 15 minutes at -80 °C.
58. Centrifuge for 30 min at 4 °C ≥ 17000 × g. Discard the supernatant.
59. Wash the pellet by adding 500 µL of 75 % ethanol without disturbing the pellet.
60. Centrifuge for 2 min at 4°C ≥17000 × g. Discard the supernatant.
61. Quick spin and remove all residual ethanol without disturbing the pellet.
62. Leave the tube opened for 5 minutes to evaporate the residual ethanol.
63. Dissolve the pellet in 20 µL of nuclease-free water and transfer to a 200-µL PCR tube. Place the tube on ice.

CRITICAL If the solution appears very viscous, dissolve the pellet in 80 µL of nuclease-free water and transfer it to a 0.5-mL PCR tube. Scale up the volumes of all subsequent steps accordingly.

Hybridization and ligation of adapters, round 1 • TIMING 30 min to overnight

CRITICAL For the adapters reannealing see Box 2. Avoid the freeze-thaw cycles for the reannealed adapters.

64. Set up the reaction by adding the following reagents sequentially to the tube containing the purified phosphorylated DNA from step 63. Mix by pipetting and perform a quick spin.

Reagent	Volume (µL)
Phosphorylated DNA (step 63)	20 µL
40 mM adapter A1 (Table 1 and Box 2)	2 µL
40 mM adapter A2 (Table 1 and Box 2)	2 µL

65. Incubate in a thermocycler programmed as outlined below:

Step	Temp	Time
Hybridization	65 °C	10 min
	16 °C	5 min

66. Take the tubes out of the thermocycler. Set up the ligation reaction by adding the following reagents sequentially to the tube:

Reagent	Volume (µL)	Final
10 × T4 ligase buffer	4 µL	1 ×
50 % PEG 8000 (w /vol)	4 µL	5 %
5 M betaine	4 µL	0.5 M
T4 DNA ligase 1 Weiss U / µL	4 µL	0.1 Weiss U / µl

CRITICAL Thaw on ice a fresh aliquot of 10 × T4 ligase buffer. Avoid freeze-thaw the aliquots.

67. Mix by pipetting, quick spin, and incubate at 16 °C in a thermocycler for 16 hours.

PAUSE POINT. The incubation can last overnight.

Streptavidin capture of biotinylated library fragments • TIMING 1 h

68. Resuspend the stock of MyOne T1 streptavidin Dynabeads by gentle vortexing.

69. Pipette 20 µl of the bead suspension into a 1.5-mL tube. Place the tube on the magnet

to capture the beads. Incubate until the liquid is clear.

70. Remove and discard the supernatant with a 200-µL filter tip.
71. Remove the tube from the magnet and add 200 µl of 1× BWT buffer, mix by pipetting.
72. Place the tube on the magnet to pellet the beads. Incubate until the liquid is clear.
73. Carefully remove and discard the supernatant with a 200-µl filter tip without disturbing the beads.
74. Repeat steps 71-73 two more times.
75. Remove the tube from the magnet and resuspend the beads in 40 µl of 2 × BWT buffer.
76. Add 40 µl of the washed bead suspension into the tube containing the ligation reaction from step 67 and mix by pipetting.
77. Incubate the tube on a rotating platform at room 15-20 rpm for 20 min at RT.

CRITICAL Ensure the beads remain in suspension during the incubation. Resuspend the beads by gently flicking the tube every 5 min. Because of the small volume, sideways rotation of the tube is preferred rather than inversion.

78. Spin the tube briefly in a microcentrifuge and place the tube on the magnet to capture the beads. Transfer the supernatant to a new 1.5-mL tube labelled “Supernatant 1” and keep it at -20 °C for the library construction quality control (Box 3)
79. Remove the tubes with the beads from the magnet, add 200 µL of 1× BWT and mix thoroughly by pipetting with a 200-µL low-binding filter tip. Transfer the entire volume to a new 1.5-mL low-binding tube.
80. Place the tube on the magnet to capture the beads. Incubate until the liquid is clear. Remove and discard the supernatant with a 200-µL tip.
81. Repeat washing steps (79-80) 2 more times with 200 µL 1 × BWT without transferring the beads to a new tube.
82. Remove the tube from the magnet and add 200 µl 1 × TE + 0.05% Tween 20 and mix by pipetting.
83. Place the tube on the magnet to pellet the beads and remove the supernatant with a 200-µL pipette tip. Repeat one more time.
84. Remove the tube from the magnet, add 200 µl of ddH_2O and mix by pipetting.
85. Place the tube on the magnet to pellet the beads and remove the supernatant with a 200-µL tip.

86. Resuspend the beads in 10 µL ddH_2O and transfer to a new 200-µL PCR tube. Place on ice and proceed immediately to the next step.

Ligation of adapters, round 2 • TIMING 4 h to overnight

87. Set up the second-round ligation reaction by adding the following reagents sequentially to the tube:

Reagent	Volume (µL)
Library bead suspension (Step 86)	10 µL
40 mM adapter A1 (Table 1 and Box 2)	1 µL
40 mM adapter A2 (Table 1 and Box 2)	1 µL

88. Mix well by pipetting and incubate in a thermocycler programmed as outlined below:

Step	Temp	Time
Hybridization	65 °C	10 min
	16 °C	5 min

89. Take the tubes out of the thermocycler. Set up the ligation reaction by adding the following reagents sequentially to the tube:

Reagent	Volume (µL)	Final
10 × T4 ligase buffer	2 µL	1 ×
50 % PEG 8000 (wt/vol)	2 µL	5 %
5 M betaine	2 µL	0.5 M
T4 DNA ligase 1 Weiss U / µL	2 µL	0.1 Weiss U / µl

90. Mix by pipetting, quick spin. Incubate at 16 °C in a thermocycler for ≥ 2 h or overnight.

91. Prepare 10 µL of fresh ligation mix by adding the following reagents sequentially in a tube on ice:

Reagent	Volume (µL)
ddH2O	7 µL
10 × T4 ligase buffer	1 µL
10 mM ATP	1 µL
T4 DNA ligase 1 Weiss U/µL	1 µL

92. Take the tube (Step 90) from the thermocycler, quick spin, and place it on the magnet to capture the beads.
93. Carefully remove 10 µL of the supernatant without disturbing the beads. Label the supernatant as "Supernatant 2" and keep it at -20 °C for the quality control of library construction (Box 3).
94. Take the tube off the magnet and add 10 µL of the fresh ligation mix (step 91). Mix by pipetting and perform a quick spin. Incubate in the thermocycler for 1 h at 16 °C.
95. Place the tube on the magnet to capture the beads. Carefully remove the supernatant without disturbing the beads.
96. Remove the tubes with the beads from the magnet, add 200 µL of 1 × BWT and mix thoroughly by pipetting with a 200-µL low-binding filter tip. Transfer the entire volume to a new 1.5-mL low-binding tube.
97. Place the tube on the magnet to capture the beads. Incubate until the liquid is clear. Remove and discard the supernatant with a 200-µL tip.
98. Repeat washing steps (96-97) 4 more times with 200 µL 1 × BWT without transferring the beads to a new tube.
99. Remove the tube from the magnet and add 200 µl 1 × TE + 0.0 5% Tween 20 and mix by pipetting.
100. Place the tube on the magnet to pellet the beads and remove the supernatant with a 200-µL pipette. Repeat one more time.
101. Remove the tube from the magnet add 200-µl of nuclease-free water and mix by pipetting.
102. Place the tube on the magnet to pellet the beads and remove the supernatant with a 200-µL tip.
103. Resuspend the beads in 20 µL of EB, transfer to a new 200-µL PCR tube and proceed to the quality control of library construction (Box 3).

PAUSE POINT The library on beads can be stored at -20 °C for up to 6 months.

Okazaki fragment library amplification • TIMING 1.5 h

104. Assemble each library amplification reaction in a low-binding 200-µL PCR tube as follows:

Component	Stock	Volume	Final
PEM1 (Table 1)	10 µM	1 µL	0.2 µM

Truseq_Index with the desired barcode (Table 1)	10 µM		0.2 µM
KAPA HiFi Fidelity Buffer	5 ×	10 µL	1 ×
Bead suspension with the bound adapter-ligated library (Step 103)		5-10 µL	
KAPA dNTP Mix	10 mM	1.5 µL	0.3 mM
Taq Kapa HiFi Hotstart Polymerase	1 U / µL	0.5 µL	0.1 U / µL
H_2O		Up to 50 µL	

105. Amplify using the following cycling protocol:

Step	Temp	Duration	Cycles
Initial denaturation	98 °C	45 sec	1
Denaturation	98 °C	15 sec	10
Annealing	60 °C	30 sec	
Extension	72 °C	30 sec	
Final extension	72 °C	1 min	1
HOLD	4 °C	∞	

CRITICAL STEP It is important to use a minimal number of amplification cycles to minimize the generation of PCR duplicates. We do not recommend exceeding 12 cycles in total. Usually, a 10-cycle library amplification synthesizes enough material for sequencing.

106. Take out the tubes from the thermocycler, quick spin, and place on the magnet to collect the beads.

107. Transfer the supernatant containing the amplified library into a new 1.5-mL low-binding tube without disturbing the beads.

108. Wash streptavidin beads once in 200 µL of EB + 0.05 % Tween 20, resuspend in 20 µL of EB, and store at -20 °C for up to several months. If necessary, Okazaki fragment library amplification (steps 104-107) can be performed one more time using the same beads as a template.

PAUSE POINT The beads can be stored at -20 °C for up to one year and the PCR product could be stored at 4 °C for 72 h or at -20 °C for up to 6 months.

Post-amplification clean-up • TIMING 1 h

109. Equilibrate AMPure XP beads at RT for at least 30 min before use.

110. Perform a 1.5 × SPRI cleanup to the supernatant that contains the amplified library (step 107) by combining the following:

Component	Volume
PCR reaction product	50 µL
AMPure XP beads	75 µL
Total volume	125 µL

111. Mix thoroughly by vortexing. Incubate the tubes at RT for 10 min to bind DNA to the beads.

112. Place the tubes on the magnet to capture the beads. Incubate until the liquid is clear. Carefully remove and discard the supernatant with a 200-µL filter tip.

113. Keeping the tubes on the magnet, add 200 µL of freshly prepared 80 % (vol/vol) ethanol. Incubate the tubes on the magnet at RT for at least 30 sec.

114. Carefully remove and discard the ethanol with a 200-µL filter tip. Repeat steps 112-113.

115. Remove all residual ethanol without disturbing the beads.

CRITICAL STEP Do not let the beads dry as it will result in irreversible DNA binding to the beads.

116. Remove the tubes from the magnet and resuspend the beads in 10.5 µL of EB.

117. Incubate the open tubes in a thermomixer for 5 min at 37 °C to elute DNA off the beads and evaporate the residual ethanol. Cover the thermomixer with a clean lid or a piece of aluminium foil to protect the tubes from dust.

118. Place the tubes on the magnet to capture the beads. Incubate until the liquid is clear.

119. Carefully transfer 10 µL of the supernatant (containing the library) to a new 1.5-mL low-binding tube without taking any beads.

PAUSE POINT The purified library could be stored at 4 °C overnight or -20 °C for up to 6 months.

Size-selection on agarose gel • TIMING 2 h

CRITICAL STEP Size selection is a critical step for optimal sequencing results.

120. Prepare a 4% Agarose gel (15 cm × 15 cm) in 1 × TAE buffer.

121. Mix 10 µL eluted DNA (step 119) with 2 µL 6 × purple gel loading dye and 1 µL SYBR Green (100 ×), and load the mix into the gel. Load DNA ladders between 20 bp and 1000 bp (like NEB low molecular weight ladder, or equivalent). Run the gel until bromophenol blue reaches ¾ of the gel length.

CRITICAL STEP The electrophoresis tank should be rinsed with deionized water in advance, and a fresh 1 × TAE buffer should be applied for electrophoresis.

CRITICAL STEP Alternatively the SYBR Green could be substituted with ethidium bromide or other dyes that do not disturb DNA migration.

Visualize the gel on a non-UV light bench and cut the bands between 150-400 bp with a clean blade.

CRITICAL STEP Do not use UV light as it damages DNA and may impact the sequencing quality.

CRITICAL STEP A visible gap should be visible between the primer dimer (128 bp) and the shortest library fragments (135-140 bp). Do not touch 128 bp band with the blade as it may lead to contamination with primer dimers.

122. Purify the DNA from the gel with the Qiagen Minelute Gel extraction kit according to the manufacturer's manual, except dissolving the agarose block at RT with gently shaking.

123. Elute DNA with 10 µL EB buffer and proceed to the quality control of the library size selection (Box 4)

PAUSE POINT The size-selected and purified library could be stored at -20 °C for up to 1 year.

? TROUBLESHOOTING

Sequencing • TIMING variable

124. Pool the libraries for multiplexing according to standard Illumina protocols.

125. Sequence the pooled libraries on an Illumina next-generation sequencing platform in single or paired-end mode. During the run set-up load the custom

sequencing primer for the read 1 (Primer R1,Table 1).

CRITICAL STEP Since the A1 adapter is shortened by 5 bp, the custom read 1 sequencing primer has to be loaded to the flowcell (following standard Illumina recommendations). Indicate to the sequencer program that a custom primer for read 1 was used before starting to run the program.

Data processing • **TIMING** variable

CRITICAL Data processing typically takes about 12 hours (tested with a classical desktop configuration: 3.5 GHz Intel Core i5 CPU with 4 cores for iMAC and 16 Go DDR4 2400 MHZ speed memory; for a dataset of ~300 million total reads).

126. Prepare/download the aligned sequencing data in .bam files.

CRITICAL The current protocol starts from the aligned data, which can be processed following standard procedures and are frequently provided by sequencing facilities. Briefly, the raw sequencing data (.fastq) need to be pre-processed into genome aligned files with the following major steps: fastqc for checking the quality of reads, cutadapt/Trim Galore/Trimmomatic for trimming adapters and low quality reads, BWA/Bowtie2 for read alignment, then Picard for marking and deleting the duplicates, samtools for sorting and indexing the aligned files.

127. Download the OKseqHMM toolkit from https://github.com/CL-CHEN-Lab/OK-Seq containing the necessary R-scripts for the following analysis steps.

CRITICAL The toolkit will count read matrices from aligned .bam files, calculate and output RFD and OEM profiles for a primary visualization (e.g. with IGV).

OkseqHMM, the R package R defines replication IZs (upward transitions of RFD profile), TZs (downward transitions of RFD profile), and two intermediate states (flat RFD profiles of low and high values (zones of leftward and rightward unidirectional replication, respectively) (Fig. 3)

Generating the output files for visualization of RFD profile and the initiation/termination zone calling by a 4-state Hidden Markov Model (HMM)

CRITICAL Besides the aligned .bam files with the corresponding indexed file (.bai), the OKseqHMM function requires the annotation coordinates for all chromosomes and their lengths.

128. Download the annotation file containing all chromosomes and their lengths from the UCSC server (e.g. hg19.chr.sizes.txt for human hg19) FTP ftp://hgdownload.cse.ucsc.edu/goldenPath/.

The program identifies automatically if the input .bam file is paired-end or single-end sequencing data, then splits the mapped reads within the .bam file into Watson (W) and Crick (C) strands, respectively, and calculates the read coverage and RFD along the reference genome. The bin size (with bin size parameter) can be defined by users depending on the data coverage anf genome size, and based on our experience, 1-kb bin size is recommended for OK-seq data of human/mouse cells, and the 50 bp bin size is recommended for yeast data.

After downloading the R scripts from GitHub, run this command line in the terminal:

source("PATH/OKseqHMM.R")

CRITICAL STEP Before executing this function, make sure that R and the necessary R packages HMM, Rsamtools, and Genomic Alignments are installed in your R working environment. Then the user can use either the command line as source ("PATH/OKseqHMM.R"), in which the PATH provides the PATH in your computer to the downloaded R package "OKseqHMM.R", or the user can load the package directly into Rstudio.

CRITICAL STEP Make sure that the chromosome coordinates within the .bam file match the ones provided in chromosome annotation file. Different sources of the reference genome having slightly different chromosomes names may cause error. (e.g. sometimes "1-22, MT" in the .bam file while the annotation file is "chr1-chr22, chrM" if you use the UCSC annotation.).

129. Run OKseqHMM with the following options:

(A) For the human data:

```
OKseqHMM(bamfile = "my.bam", thresh = 10, chrsizes = "hg19.chr.size.txt", binSize=1000, winS=15, fileOut = "my_hmm"))
```

(B) For the yeast data:

```
OKseqHMM(bamfile = "my.bam", thresh = 10, chrsizes = " sacCer3.chrom.sizes.txt", binSize=50, winS=15, fileOut = "my_hmm"))
```

CRITICAL STEP Bin size may need to be adjusted relative to the genome size of the analyzed species and the coverages of your data.

CRITICAL STEP

"My.bam" is your input path of .bam file;

"thresh" is the threshold to eliminate the low read coverage bins;

"chrsizes" is your path linked with the annotation file containing the length of each chromosome;

"binSize" is the adjacent bin size in bp to calculate the read coverage and RFD;

"winS" is the smoothing window size for the HMM calling;

"fileOut" is the path of storage as well as the prefix of name for your output files.

? TROUBLESHOOTING

130. The function OKseqHMM will generate automatically a series of output files including:

(1-4) Two .bam files, and their corresponding index .bai files, for the reads generated from the Watson and Crick strands, respectively.

(5-6) Two bedgraph files containing RFD values in the adjacent bins of defined size. ("_RFD.bedgraph"), with the raw RFD values and the smoothed values in adjacent windows (span window size) RFD with the defined bin size and the smoothing one with span window size)

(7) log file ("_log.txt") that records all of the parameters you use and also the default setting information.

(8) HMM result in a text file ("_HMM.txt") that records all of the global optimal hidden states calculated by HMM Viterbi algorithm.

(9) HMM result in a text file ("_HMMpropa.txt") that records all of the previous state positions that caused the maximum local probability of a state by HMM posterior algorithm.

(10-17) 8 text files recording the genomic positions (.bed) and the corresponding probabilities (.txt) for the final identified optimal states:

"_HMMsegments_IZ.bed/txt" is for the replication initiation zone calling result.

"_HMMsegments_TZ. bed/txt" is for the replication termination zone calling result.

"_HMMsegments_highFlatZone. bed/txt" and "_HMMsegments_LowFlatZone. bed/txt" are the results of two intermediate flat states.

CRITICAL RFD Bedgraph files can be visualized directly in genomic browsers, *e.g.* IGV[86].

CRITICAL You can also further transform the bedgraphs into bigwig by the UCSC tool bedGraphToBigWig (http://hgdownload.soe.ucsc.edu/admin/exe/) to get binary compressed files.

CRITICAL Additional details about the parameters are listed in Box 5.

Generating the output files for visualization of the RFD transitions

CRITICAL OKseqOEM function allows investigating the local origin efficiency metrics (i.e. deltaRFD) [30] at multiple scales.

131. Download the R scripts from GitHub, run this command line in the terminal: source("PATH/OKseqOEM.R") with the following options:

(A) For the human data:

```
OKseqOEM(bamInF="path_to_bam_Forward_strand",bamInR="path_to_bam_Reverse_strand",chrsizes="hg19.chr.size.txt",fileOut="path/name_of_my_OEM",binSize=1000,binList=c(1,10,20,50,100,250,500,1000))
```

(B) For the yeast data:

```
OKseqOEM(bamInF="path_to_bam_Forward_strand",bamInR="path_to_bam_Reverse_strand",chrsizes="sacCer3.chrom.sizes.txt", fileOut="path/name_of_my_OEM", binSize=50, binList=c(1,20,100,200,300,400,500))
```

CRITICAL STEP

"bamInF" and "bamInR" are the paths to the two .bam files of Watson and Crick strand respectively, generated by OKseqHMM function;

"chrsizes" is the path to annotation coordinates containing chromosome length information;

"fileOut" is the path of storage as well as the prefix of the name given by the user to be used (e.g. ~/Desktop/Okseq_results/my_HMM) for the output file,

"binsize" is to define the adjacent bin size in bp to calculate the read coverage for RFD,

"binList" is to define a series of window sizes as different visualization scales that you would like to output the OEM results (e.g. for yeast cells, you will get 50 bp, 1 kb, 5 kb, 10 kb, 15 kb, 20 kb, 25 kb window scales OEM files if you set binsize = 50 and binList = c(1, 20, 100, 200, 300, 400, 500) as indicated in the previous given common line).

? TROUBLESHOOTING

132. The OKseqOEM function will generate automatically a series of wiggle (.wig) files calculated by using different sliding window sizes defined by "binList".

133. Convert wiggle to bigwig format by using the UCSC tool wigToBigWig (http://hgdownload.soe.ucsc.edu/admin/exe/) for the visualization.

Generating the output files for the generation of the average profile and heatmap of RFD values around the regions of interest

The shell-based script "average_profile_heatmap.sh" shows us the template on how to use the deepTools to generate the average profile and heatmap around or among the regions of interest (such as around the transcription start sites, transcription termination sites, within the annotated genes, around the initiation zones, etc.) by using the "computeMatrix" and "plotProfile"/ "plotHeatmap" functions, defining the two upstream and downstream border sizes and intragenic body size and also the other parameters indicated in the script.

134. Compute the matrix of values by running the command line in terminal or Rstudio:

computeMatrix scale-regions --regionsFileName {your bed file of interested regions/genes PATH e.g.codingGenes.bed} --beforeRegionStartLength {e.g. 10000} --afterRegionStartLength {e.g. 10000} --regionBodyLength {e.g. 20000} --binSize {e.g. 1000} --scoreFileName {RFD bigwig file PATH e.g. Hela.EdC.Combined_OkaSeq.RFD.bw} --outFileName {e.g. "OUTPUT.matrix"} --missingDataAsZero –skipZeros

135. For obtaining average profile run "plotProfile" function as follows:

plotProfile --matrixFile {e.g. "OUTPUT.matrix"} --outFileName {e.g. "RFD_averageProfile.stGeneLength.png"} --averageType mean --startLabel {e.g. start/TSS} --endLabel {e.g. end/TTS} --plotType se

136. For obtaining the average profile and the heatmap, proceed following this example to plot the OEM around the centre of IZ with the extension of +/-100 kb in different scales (from 1 kb to 1 Mb) and the bigwig files used in the example can be found at https://github.com/CL-CHEN-Lab/OK-Seq/tree/master/published_results/HeLa:

computeMatrix reference-point --regionsFileName {your IZ bed file PATH e.g. HeLa_hmm_HMMsegments_IZ.bed} --beforeRegionStartLength {e.g. 100000} --afterRegionStartLength {e.g. 100000} --binSize {e.g. 1000} --scoreFileName {series of OEM bigwig file PATH e.g. 20130819CGM130726.Hela_OEM_10kb.bw 20130819CGM130726.Hela_OEM_20kb.bw

20130819CGM130726.Hela_OEM_50kb.bw
20130819CGM130726.Hela_OEM_100kb.bw
20130819CGM130726.Hela_OEM_250kb.bw
20130819CGM130726.Hela_OEM_500kb.bw
20130819CGM130726.Hela_OEM_1Mb.bw} --outFileName {e.g. "OUTPUT.matrix"} --missingDataAsZero --skipZeros --referencePoint center

137. To plot the profile and heatmap, use the matrix calculated by "computeMatrix", run "plotHeatmap" to get the figure output:

plotHeatmap --matrixFile {e.g. "OUTPUT.matrix"} --outFileName {e.g. "OEM_sortbyLength.png"} --whatToShow "plot, heatmap and colorbar" --refPointLabel center --samplesLabel {e.g. "HeLa 10kb" "HeLa 20kb" "HeLa 50kb" "HeLa 100kb" "HeLa 250kb" "HeLa 500kb" "HeLa 1Mb"} --sortUsing region_length --sortRegions ascend

CRITICAL STEP Make sure that you already installed the deepTools and the python environment (the recommended version is python 3.6.4 or less. The latest python version could cause some incompatibility issues with deepTools [85]. Refer to the DeepTools manual for different functions and set up the parameters (https://deeptools.readthedocs.io/en/develop/index.html).

• TIMING

Step 1, Cell culture, EdU labelling, and cell harvesting: 2-7 days of cell culture, 2 hours of labelling and harvesting

Steps 2-21, Extraction of the genomic DNA: 2 hours with overnight incubation

Steps 22-40, Size-fractionation of denatured genomic DNA on neutral sucrose gradients: 3.5 h of handling and 17 h of centrifugation

Steps 41-48, Click biotinylation: 2 h

Steps 49-53, RNA hydrolysis: 20 min

Steps 54-63, DNA phosphorylation and precipitation: 1.5 h

Steps 64-67, Hybridization and ligation of adapters, round 1: 30 min to overnight

Steps 68-86, Streptavidin capture of biotinylated library fragments: 1 h

Steps 87-103, Hybridization and ligation of adapters, round 2: 4 h to overnight

Steps 104-108, Okazaki fragment library amplification: 1.5 h

Steps 109-119, Post-amplification clean-up: 1 h

Steps120-124, Library size-selection: 2 h

Steps 125-126, Sequencing: variable

Steps 127-128, Data processing: variable

Steps 129-131, Generating the output files for visualization of RFD profile and the initiation/termination zone calling by Hidden Markov Model (HMM): variable

Steps 132-134, Generating the output files for visualization of the RFD transitions: variable

Steps 135-138, Generating the output files for the generation of the average profile and heatmap of RFD values around the regions of interest: variable

TROUBLESHOOTING

Please find the advice for troubleshooting in Table 2.

ANTICIPATED RESULTS

DNA size-fractionation

Genomic DNA preparation from 3-10 x 10^8 human cells typically yields 2-3 mg DNA, which is then denatured and size-fractionated on 4-6 × sucrose gradients. When visualizing the DNA in each 1-mL fraction (Box 1), the DNA size linearly increases in the fractions from top to bottom (Box 1). Typically, Okazaki fragments (< 200 nt) are present in the top 1mL fractions 1 - 8. It is important to avoid contamination from the lower fractions containing high molecular weight labelled nascent replicated strands.

Library size distribution

The library fragment size should range from 150 to 300 bp. To evaluate if the library preparation is successful a PCR control can be performed (Box 3). A smear >140 bp containing the library with inserts should be more prominent than the adapter dimer (at 128 bp) (Box 3). After gel size selection, ideally no or very few adapter dimers should be present (Box 4). If the dimer peak is more abundant than the smear this is an indication of a low-complexity library, which will require repeating the size-selection step and may impact the data quality.

Sequencing results

The examples of sequencing results of OK-seq in yeast and human are shown in Fig 3. Replication fork directionality (RFD) profiles are calculated based on the proportion of the read counts from the Crick and Watson genomic strands and reflect the locus-specific average fork direction (Fig. 3). HMM detection of RFD transitions detects the initiation and termination zones. Automated approach OKseqHMM efficiently detects site-specific (yeasts) and broad zones (human cells) of replication initiation events and the regions of predominantly unidirectional fork movement (flat segments). Applying OKseqOEM allows to assess local initiation efficiency at different scales (Fig. 3).

REPORTING SUMMARY:

Further information on research design is available in the Nature Research Reporting Summary linked to this article

DATA AVAILABILITY

Published available sequencing raw and processed datasets analyzed in this work are available in SRA: SRP065949 (human HeLa cells) and ENA: PRJEB36782 (yeasts).

CODE AVAILABILITY

The bioinformatics tool and all example datasets underlying this paper are available at the following GitHub page: https://github.com/CL-CHEN-Lab/OK-Seq.

BOXES and TABLES

Box 1 Quality control of DNA size fractionation • TIMING 1 h

1) Mix 10 µL of each gradient fraction from 2 to 10 with 10 µL Gel loading buffer II in a 1.5 m- tube.
2) Heat the tubes at 94 °C for 5 min.
3) Chill the tubes on ice for 5 min.
4) Set up a TBE-Urea gel (10 %, 1 mm) on the vertical electrophoresis system with 1 × TBE buffer. Flush carefully each well with 1 × TBE buffer.
5) Prewarm the gel by running empty for 10 minutes at 400 V.

6) Quick spin the samples and load the entire volume to the wells.
7) Run at 180 V until the bromophenol blue reaches the bottom of the gel (usually 30-40 min).

CRITICAL STEP For the homemade gel, run the gel at 400 V until the bromophenol blue approaches the bottom (usually 10-12 min).

8) Stain the gel by immersing in 20 mL of freshly prepared 1 × Sybr Gold stain in TBE.
9) Visualize at a UV transilluminator.
10) Determine the gradient fractions containing the fragments of interest (≤ 250 nt).

CRITICAL The DNA size is increasing in the fractions from top to bottom. The tRNA and 5S rRNA serves as internal size markers. Typically, the 1-mL fractions 1 to 8 are combined to collect Okazaki fragments.

CRITICAL STEP The quality control of gradient fractionation may also be performed using 3 % neutral agarose gels.

Box 1 Figure legend: Quality control for DNA size-fractionation. Representative electrophoresis in 10 % Urea-PAGE. 2-10, 2^{nd} to 10^{th} 1-mL gradient fractions; LMW, NEB low molecular weight marker.

Box 2 Adapter preparation

CRITICAL STEP To obtain double-stranded adapters A1 and A2 with single-stranded random hexamer overhangs, anneal the Adapter oligonucleotide "top" with the Adapter oligonucleotide "bottom".

1) Dissolve the adapter oligomers (Table 1) to 100 µM with nuclease-free H_2O and vortex to achieve complete dissolution.
2) Prepare two 200-µL PCR tubes labelled A1 and A2 for Adapter 1 and Adapter 2 respectively.
3) Assemble each adapter reannealing reaction in a PCR tube on ice by adding in the following order:

Reagent	A1	A2	Volume (µL)
Top strand 100 µM	$A1_{top}$	$A2_{top}$	20 µL
Bottom strand 100 µM	$A1_{bottom}$	$A2_{bottom}$	20 µL

NaCl 5M			0.5 µL
Water			9.5 µL

4) Mix well by pipetting, quick spin hybridization reaction in a thermal cycler: cool down from 94 °C to 16 °C at 0.1 °C / s.
5) Chill on ice and aliquot the annealed adapters into 5 µL.

CRITICAL STEP Keep at -20 °C for up to 6 months. Avoid thaw-freezing to preserve the phosphorylation modifications on the oligomers.

Box 3 Quality control of library construction

1) Assemble 4 amplification reactions in 4 PCR tubes on ice as follows:

Component	Stock	Volume	Final
PEM1 (Table 1)	10 µM	0.2 µL	0.1 µM
Truseq_Index (Table 1)	10 µM	0.2 µL	0.1 µM
Taq DNA polymerase Buffer	10 ×	2 µL	1 ×
Template		1 µL	
dNTP Mix	10 mM	0.4 µL	0.2 mM
Taq DNA polymerase	5 U / µL	0.2 µL	0.05 U / µL
H_2O		Up to 20 µL	

2) Add 1 µL of the following templates to each PCR reaction tube: 1 - 1µl of nuclease-free H_2O (negative control); 2 - 1 µL of the bead suspension with bound adapter-ligated library (Step 103); tube 3 - 0.2 µL of ligation supernatant 1 (step 80) plus 0.8 µL nuclease-free H_2O; tube 4 - 1 µL ligation supernatant 2 (step 93).
3) Amplify using the following cycling protocol:

Step	Temp	Duration	Cycles
Initial denaturation	98 °C	45 sec	1
Denaturation	98 °C	15 sec	25-30
Annealing	60 °C	30 sec	
Extension	72 °C	30 sec	
Final extension	72 °C	1 min	1
HOLD	4 °C	∞	

4) Prepare a 10% TBE PAGE gel
5) Mix 10 µL of PCR product (Step 3) with 2 µL of 6 ×purple loading dye, and load the mix into the gel. Run the gel until bromophenol blue reaches the bottom of the gel.
6) Stain the gel by immersing in 20 mL of freshly prepared 1 × SybrGold for 5 min.
7) Visualize at a UV transilluminator and compare the lanes.

CRITICAL In the PCR reaction run with the bead-bound adapter-ligated library (lane 2), the 128 bp band corresponds to the self-ligated adapter dimers and the smear above contains the library with inserts. As an indicator of a successful library, the dimer band has to be visible but less prominent than the library smear. In PCR reactions run with the supernatants 1 and 2, no or very little smear above 128 bp is observed (lanes 3 and 4).

Box 3 Figure legend: Quality control for the library construction. Representative electrophoresis in 10 % TBE-PAGE. "LMW" - NEB low molecular weight marker. "H_2O"- PCR reaction run without template (negative control). "Beads" - PCR reaction run with the bead-bound library. "Sup 1" and "Sup2" - PCR reactions run with supernatants 1 and 2.

Box 4 Quality control of the library size-selection

1) Measure the library concentration of the size-selected and purified libraries using a Qubit dsDNA HS Kit following the manufacturer's recommendations.
2) Check the fragment size distribution by running 1 µL on an Agilent Bioanalyzer High Sensitivity DNA Chip. A typical size-selected library ranges between 145 bp and 250 bp.

Box 4 Figure legend: Quality control for library size-selection Representative profile of OK-seq libraries obtained by Agilent Bioanalyzer. An average library size of 145-250 bp is desired.

? TROUBLESHOOTING

Box 5 Additional parameters of the OKseqHMM toolkit

To run the OKseqHMM function, one needs to pre-define the initial start probabilities for the 4 states of HMM, including the transition matrix containing the probabilities that the system goes from one state to another, the emission probability matrix between states and observations, and the 5 quantiles of RFD as following:

st=c("D", "L", "H", "U"), sym=c("V", "W", "X", "Y", "Z"), pstart=rep(1/4, 4),

```
pem=t(matrix(c(0.383886256, 0.255924171, 0.170616114, 0.113744076,
0.075829384,
                .10,.20,.40,.20,.10,
                .10,.20,.40,.20,.10,
                0.022222222, 0.033333333, 0.066666667, 0.211111111,
0.666666667),
                    ncol=4)),
ptrans=t(matrix(c(0.9999,0.000020,0,0.000080,
                  0,0.999,0,0.001,
                  0.001,0,0.999,0,
                  0.000080,0,0.000020,0.9999),
                    ncol=4)).
quant=c(-1, -0.0082058939609862, -0.00141890249101162, 0.00103088286465956,
0.00800467305420799, 1))
```

These parameters and probabilities were validated with the OK-seq dataset of HeLa cells [28]. We have successfully applied them to different human, mouse, and yeast OK-seq datasets, which all got satisfactory results with these pre-setting parameters. However, users could modify these parameters to optimize the results for their dataset.

Table 1. Oligonucleotides used in the study

Oligo name	Sequences (5' to 3')
$A1_{top}$	ACACTCTTTCCCTACACGACGCTCTTCC
$A1_{bottom}$	NNNNNNGGAAGAGCGTCGTGTAGGGAAAGAGTG
$A2_{top}$	[Phos]-AGATCGGAAGAGCACACGTCTGAACTCCAGTCA[ddC]
$A2_{bottom}$	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNN[ddC]
PEM_1.0	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACG ACGCTCTTCC
TruSeq_Index 1	CAAGCAGAAGACGGCATACGAGATcgtgatGTGACTGGAGTTCAG ACGTGTGCTCTTCCGATCT
TruSeq_Index 2	CAAGCAGAAGACGGCATACGAGATacatcgGTGACTGGAGTTCAG ACGTGTGCTCTTCCGATCT

TruSeq_Index 3	CAAGCAGAAGACGGCATACGAGATgcctaaGTGACTGGAGTTCAG ACGTGTGCTCTTCCGATCT
TruSeq_Index 4	CAAGCAGAAGACGGCATACGAGATtggtcaGTGACTGGAGTTCAG ACGTGTGCTCTTCCGATCT
TruSeq_Index 5	CAAGCAGAAGACGGCATACGAGATcactgtGTGACTGGAGTTCAG ACGTGTGCTCTTCCGATCT
TruSeq _Index 6	CAAGCAGAAGACGGCATACGAGATattggcGTGACTGGAGTTCAG ACGTGTGCTCTTCCGATCT
TruSeq_Index 7	CAAGCAGAAGACGGCATACGAGATtcaagtGTGACTGGAGTTCAG ACGTGTGCTCTTCCGATCT
TruSeq_Index 8	CAAGCAGAAGACGGCATACGAGATctgatcGTGACTGGAGTTCAG ACGTGTGCTCTTCCGATCT
Primer R1	ACACTCTTTCCCTACACGACGCTCTTCC

Table 2. Troubleshooting

Steps	Problems	Possible reasons	Solutions
6	Nonhomogeneous / non-transparent solution	Cells aggregation formed before cell lysis, and/or inadequate proteinase K treatment.	Thoroughly resuspend the cells before adding SDS. Add additional proteinase K to 0.1 mg/mL, invert gently to mix well, and incubate at 42 °C for an additional 2 h.
23	Incomplete DNA dissolution	Ethanol residues and/or insufficient dissolution time.	Incubate opened tubes with DNA solution at 37 °C for1 h. Carefully resuspend with a wide-bore tip.
40	Final volume > 375 μL	Insufficient centrifugation or/and presence of polysaccharides.	• Spin for an additional 10 min in step 39. • Scale up the reagents in the following steps.
124	Prominent adapter-dimer peak	Low-complexity library and/or insufficient gel size selection.	• Amplify the library again with the beads from step 108. • Perform a double-size selection of the library with Ampure

			beads, ratio 1:1.25 (if the total library amount is >10 ng).
124	Smear containing libraries is absent or very weak	An insufficient number of starting cells.	• Increase starting cell number • Ensure the cell are EdU-labelled using flow cytometry. Check the fraction of cells in S-phase and EdU-positive cells). • For cell lines or conditions having less than 20% of cells in the S-phase, increase the number of starting cells. • As a control, perform OK-seq on a well-proliferating cell line in parallel (HeLa).
130	Function execution interrupted by error	• The pre-required R packages are not installed. • Incomplete parameters Inappropriate 'thresh' • Different annotations are used in the aligned files and 'chrsizes'. • Inappropriate 'binSize'.	• Install all R packages and make sure all input fields are filled before execution. • Check whether the chromosome names in your aligned files are consistent with your input annotation. • Set a smaller 'thresh' or a larger 'binSize' if the sequencing depth is low.
132	Function execution was interrupted by error	• Incomplete parameters • Inappropriate 'binSize' and 'binList'	• Complete all the required fields before execution. • Modify the values of 'binSize' and test the scales of 'binList' based on the data.

FIGURE LEGENDS

Fig. 1: Detection of replication initiations and terminations by OK-seq. (**a**) Okazaki fragment strandedness indicates the direction of ongoing replication forks. Watson strand Okazaki fragments (red) are generated from leftward-oriented forks. Crick strand Okazaki fragments (blue) are generated from rightward-oriented forks. RFD, the population-averaged fork directionality is computed as a proportion of Okazaki fragment reads from Crick and Watson strands. (**b**) The RFD profile reflects the location, nature and efficiency of replication initiation. Site-specific initiation (left and centre panel) results in an abrupt positive shift of RFD whereas initiation zone results in a progressive positive shift of RFD (right panel) (IZ). The amplitude of the RFD shift reflects the initiation efficiency. (**c**) Negative shifts of RFD reflect the sites and zones of fork merging (predominantly termination zones).

Fig. 2: Experimental workflow and data processing pipelines of OK-seq. (**a**) Illustration of the key experimental steps. Unreplicated DNA is in black and two replicated DNA strands are in red and in blue. Watson and Crick strand Okazaki fragments are shown as red and blue arrows; EdU (green dots), biotin (red dots), Streptavidin magnetic beads (black), and double-stranded adapters (grey and yellow). (**b**) Flowchart representing data analysis pipeline. OKseqHMM allows to split Watson and Crick strand reads and to compute the RFD values at defined bin size. Further, the automated detection of zones of replication initiation, termination and unidirectional fork movement is achieved by segmentation of the RFD profile into upward, downward and flat segments by HMM. OKseqOEM tool computes OEM at different genomic scales. Average plot allows creating the heatmaps and linear plots to explore RFD patterns around genomic features of interest.

Fig. 3 Representative results for OK-seq Okazaki fragment Watson stand (red) and Crick strand (blue) read counts, RFD computed in 1 kb windows and OEM at indicated scales. Initiation zones (yellow) and termination zones (teal blue), flat segments of unidirectional replication (pink), detected by OKseqHMM. Panels **a** show data for HeLa cells[28] and panel **b** shows data for yeast *S. cerevisiae* [31].

ACKNOWLEDGMENTS

X.W. is supported by The Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 31900415). Y.L. thanks Agence Nationale pour la Recherche (ANR) for

providing her Ph. D. fellowship. C.T., Y.D.-C., C-L.C, O.H and N.P. thank the ANR grant BLAN2010–161501 (REFOPOL). Work in the O.H. lab is supported by the ANR grants 18-CE45-0002 (NanoPoRep) and 19-CE12-0028 (HUDROR). Work in the C-L.C lab is supported by the YPI program of I. Curie, the ATIP-Avenir program from Centre national de la recherche scientifique (CNRS) and Plan Cancer (grant number ATIP/AVENIR: N°18CT014-00); ANR grant 19-CE12-0016-02 (ReDeFINe) and 19-CE12-0020-02 (TELOCHROM); and Institut National du Cancer (INCa) grant PLBIO19-076. N.P. is the recipient of the CNRS-INSERM ATIP-Avenir grant and YPI funding from Institute Gustave Roussy; and is supported by LabEx "Who Am I?" #ANR-11-LABX-0071; the Université de Paris IdEx #ANR-18-IDEX-0001 and ANR grant 19-CE12-0030-01 (INTEGER).

AUTHOR CONTRIBUTIONS

O.H., C-L.C. and N.P. conceived and supervised the project. N.P. developed the OK-seq method in mammalian cells; X.W. adapted the method for yeast cells. Y.L. Y.D-C., C.T. and C-L.C. developed the bioinformatical approach and built the analysis pipeline. X.W., Y.L., O.H., C-L.C., and N.P. wrote the manuscript with input from all authors.

COMPETING INTERESTS

Authors declare no competing interests

SUPPLEMENTARY INFORMATION

Supplementary method– OK-seq in *S. cerevisiae*

Related links

Key references using this protocol

1. Petryk, N. et al. Replication landscape of the human genome. Nature communications 7, 10208, doi:10.1038/ncomms10208 (2016)
2. Wu, X. et al. Developmental and cancer-associated plasticity of DNA replication preferentially targets GC-poor, lowly expressed and late-replicating regions. Nucleic acids research 46, 10157-10172, doi:10.1093/nar/gky797 (2018)

3. Hennion, M. et al. FORK-seq: replication landscape of the Saccharomyces cerevisiae genome by nanopore sequencing. Genome Biol 21, 125, doi:10.1186/s13059-020-02013-3 (2020).

REFERENCES

1 Huberman, J. A. & Riggs, A. D. On the mechanism of DNA replication in mammalian chromosomes. *J Mol Biol* **32**, 327-341, doi:10.1016/0022-2836(68)90013-2 (1968).

2 Hamlin, J. L., Mesner, L. D. & Dijkwel, P. A. A winding road to origin discovery. *Chromosome Res* **18**, 45-61, doi:10.1007/s10577-009-9089-z (2010).

3 Hyrien, O. Peaks cloaked in the mist: The landscape of mammalian replication origins. *The Journal of cell biology* **208**, 147-160, doi:10.1083/jcb.201407004 (2015).

4 Hulke, M. L., Massey, D. J. & Koren, A. Genomic methods for measuring DNA replication dynamics. *Chromosome Res*, doi:10.1007/s10577-019-09624-y (2019).

5 Lebofsky, R., Heilig, R., Sonnleitner, M., Weissenbach, J. & Bensimon, A. DNA replication origin interference increases the spacing between initiation events in human cells. *Mol Biol Cell* **17**, 5337-5345, doi:10.1091/mbc.E06-04-0298 (2006).

6 Demczuk, A. *et al.* Regulation of DNA replication within the immunoglobulin heavy-chain locus during B cell commitment. *PLoS biology* **10**, e1001360, doi:10.1371/journal.pbio.1001360 (2012).

7 Anglana, M., Apiou, F., Bensimon, A. & Debatisse, M. Dynamics of DNA replication in mammalian somatic cells: nucleotide pool modulates origin choice and interorigin spacing. *Cell* **114**, 385-394, doi:10.1016/s0092-8674(03)00569-5 (2003).

8 Cadoret, J. C. *et al.* Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 15837-15842, doi:10.1073/pnas.0805208105 (2008).

9 Besnard, E. *et al.* Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nature structural & molecular biology* **19**, 837-844, doi:10.1038/nsmb.2339 (2012).

10 Karnani, N., Taylor, C. M., Malhotra, A. & Dutta, A. Genomic study of replication initiation in human chromosomes reveals the influence of transcription regulation and chromatin structure on origin selection. *Mol Biol Cell* **21**, 393-404, doi:10.1091/mbc.E09-08-0707 (2010).

11 Mukhopadhyay, R. *et al.* Allele-specific genome-wide profiling in human primary erythroblasts reveal replication program organization. *PLoS genetics* **10**, e1004319, doi:10.1371/journal.pgen.1004319 (2014).

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

13 Mesner, L. D. *et al.* Bubble-chip analysis of human origin distributions demonstrates on a genomic scale significant clustering into zones and significant association with transcription. *Genome research* **21**, 377-389, doi:10.1101/gr.111328.110 (2011).

14 Mesner, L. D. *et al.* Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early- and late-firing origins. *Genome research*, doi:10.1101/gr.155218.113 (2013).
15 Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 139-144, doi:10.1073/pnas.0912402107 (2010).
16 Chen, C. L. *et al.* Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome research* **20**, 447-457, doi:10.1101/gr.098947.109 (2010).
17 Zhao, P. A., Sasaki, T. & Gilbert, D. M. High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome Biol* **21**, 76, doi:10.1186/s13059-020-01983-8 (2020).
18 Koren, A. *et al.* Genetic Variation in Human DNA Replication Timing. *Cell*, doi:10.1016/j.cell.2014.10.025 (2014).
19 Hulke, M. L., Massey, D. J. & Koren, A. Genomic methods for measuring DNA replication dynamics. *Chromosome Res* **28**, 49-67, doi:10.1007/s10577-019-09624-y (2020).
20 Lobry, J. R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* **13**, 660-665, doi:10.1093/oxfordjournals.molbev.a025626 (1996).
21 Touchon, M. *et al.* Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 9836-9841, doi:10.1073/pnas.0500577102 (2005).
22 Huvet, M. *et al.* Human gene organization driven by the coordination of replication and transcription. *Genome research* **17**, 1278-1285, doi:10.1101/gr.6533407 (2007).
23 Chen, C. L. *et al.* Replication-associated mutational asymmetry in the human genome. *Mol Biol Evol* **28**, 2327-2337, doi:10.1093/molbev/msr056 (2011).
24 Audit, B. *et al.* Open chromatin encoded in DNA sequence is the signature of 'master' replication origins in human cells. *Nucleic Acids Res* **37**, 6064-6075, doi:10.1093/nar/gkp631 (2009).
25 Guilbaud, G. *et al.* Evidence for sequential and increasing activation of replication origins along replication timing gradients in the human genome. *PLoS computational biology* **7**, e1002322, doi:10.1371/journal.pcbi.1002322 (2011).
26 Baker, A. *et al.* Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines. *PLoS computational biology* **8**, e1002443, doi:10.1371/journal.pcbi.1002443 (2012).
27 Green, P., Ewing, B., Miller, W., Thomas, P. J. & Green, E. D. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet* **33**, 514-517, doi:10.1038/ng1103 (2003).
28 Petryk, N. *et al.* Replication landscape of the human genome. *Nat Commun* **7**, 10208, doi:10.1038/ncomms10208 (2016).
29 Smith, D. J. & Whitehouse, I. Intrinsic coupling of lagging-strand synthesis to chromatin assembly. *Nature* **483**, 434-438, doi:10.1038/nature10895 (2012).
30 McGuffee, S. R., Smith, D. J. & Whitehouse, I. Quantitative, genome-wide analysis of eukaryotic replication initiation and termination. *Molecular cell* **50**, 123-135, doi:10.1016/j.molcel.2013.03.004 (2013).
31 Hennion, M. *et al.* FORK-seq: replication landscape of the Saccharomyces cerevisiae genome by nanopore sequencing. *Genome Biol* **21**, 125, doi:10.1186/s13059-020-02013-3 (2020).

32 Liu, Y., Wu, X., D'aubenton-Carafa, Y., Thermes, C. & Chen, C.-L. OKseqHMM: a genome-wide replication fork directionality analysis toolkit. *bioRxiv*, 2022.2001.2012.476022, doi:10.1101/2022.01.12.476022 (2022).
33 Blin, M. *et al.* DNA molecular combing-based replication fork directionality profiling. *Nucleic Acids Res* **49**, e69, doi:10.1093/nar/gkab219 (2021).
34 Wang, W. *et al.* Genome-wide mapping of human DNA replication by optical replication mapping supports a stochastic model of eukaryotic replication. *Molecular cell* **81**, 2975-2988.e2976, doi:10.1016/j.molcel.2021.05.024 (2021).
35 Wu, X. *et al.* Developmental and cancer-associated plasticity of DNA replication preferentially targets GC-poor, lowly expressed and late-replicating regions. *Nucleic Acids Res* **46**, 10157-10172, doi:10.1093/nar/gky797 (2018).
36 Petryk, N. *et al.* MCM2 promotes symmetric inheritance of modified histones during DNA replication. *Science* **361**, 1389-1392, doi:10.1126/science.aau0294 (2018).
37 Chen, Y. H. *et al.* Transcription shapes DNA replication initiation and termination in human cells. *Nature structural & molecular biology* **26**, 67-77, doi:10.1038/s41594-018-0171-0 (2019).
38 Li, Z. *et al.* DNA polymerase alpha interacts with H3-H4 and facilitates the transfer of parental histones to lagging strands. *Sci Adv* **6**, eabb5820, doi:10.1126/sciadv.abb5820 (2020).
39 Tubbs, A. *et al.* Dual Roles of Poly(dA:dT) Tracts in Replication Initiation and Fork Collapse. *Cell* **174**, 1127-1142.e1119, doi:10.1016/j.cell.2018.07.011 (2018).
40 Kirstein, N. *et al.* Human ORC/MCM density is low in active genes and correlates with replication time but does not delimit initiation zones. *Elife* **10**, doi:10.7554/eLife.62161 (2021).
41 Hyrien, O. M., C.; Mechali, M. TransitioninSpecificationofEmbryonic Metazoan DNA ReplicationOrigins. *Science* **270**, 994-997 (1995).
42 Dijkwel, P. A., Wang, S. & Hamlin, J. L. Initiation sites are distributed at frequent intervals in the Chinese hamster dihydrofolate reductase origin of replication but are used with very different efficiencies. *Mol Cell Biol* **22**, 3053-3065, doi:10.1128/mcb.22.9.3053-3065.2002 (2002).
43 Powell, S. K. *et al.* Dynamic loading and redistribution of the Mcm2-7 helicase complex through the cell cycle. *EMBO J* **34**, 531-543, doi:10.15252/embj.201488307 (2015).
44 Gros, J. *et al.* Post-licensing Specification of Eukaryotic Replication Origins by Facilitated Mcm2-7 Sliding along DNA. *Molecular cell* **60**, 797-807, doi:10.1016/j.molcel.2015.10.022 (2015).
45 Promonet, A. *et al.* Topoisomerase 1 prevents replication stress at R-loop-enriched transcription termination sites. *Nat Commun* **11**, 3940, doi:10.1038/s41467-020-17858-2 (2020).
46 Brison, O. *et al.* Transcription-mediated organization of the replication initiation program across large genes sets common fragile sites genome-wide. *Nat Commun* **10**, 5693, doi:10.1038/s41467-019-13674-5 (2019).
47 Letessier, A. *et al.* Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. *Nature* **470**, 120-123, doi:10.1038/nature09745 (2011).
48 Le Tallec, B. *et al.* Common fragile site profiling in epithelial and erythroid cells reveals that most recurrent cancer deletions lie in fragile sites hosting large genes. *Cell Rep* **4**, 420-428, doi:10.1016/j.celrep.2013.07.003 (2013).
49 Hamperl, S., Bocek, M. J., Saldivar, J. C., Swigut, T. & Cimprich, K. A. Transcription-Replication Conflict Orientation Modulates R-Loop Levels and

Activates Distinct DNA Damage Responses. *Cell* **170**, 774-786.e719, doi:10.1016/j.cell.2017.07.043 (2017).
50 Manzo, S. G. *et al.* DNA Topoisomerase I differentially modulates R-loops across the human genome. *Genome Biol* **19**, 100, doi:10.1186/s13059-018-1478-1 (2018).
51 Park, K. *et al.* Aicardi-Goutières syndrome-associated gene SAMHD1 preserves genome integrity by preventing R-loop formation at transcription-replication conflict regions. *PLoS genetics* **17**, e1009523, doi:10.1371/journal.pgen.1009523 (2021).
52 Bayona-Feliu, A., Barroso, S., Muñoz, S. & Aguilera, A. The SWI/SNF chromatin remodeling complex helps resolve R-loop-mediated transcription-replication conflicts. *Nat Genet* **53**, 1050-1063, doi:10.1038/s41588-021-00867-2 (2021).
53 Andrianova, M. A., Bazykin, G. A., Nikolaev, S. I. & Seplyarskiy, V. B. Human mismatch repair system balances mutation rates between strands by removing more mismatches from the lagging strand. *Genome research* **27**, 1336-1343, doi:10.1101/gr.219915.116 (2017).
54 Jaksik, R., Wheeler, D. A. & Kimmel, M. Detection and characterization of replication origins defined by DNA polymerase epsilon. *bioRxiv*, 2021.2007.2027.453931, doi:10.1101/2021.07.27.453931 (2021).
55 Shi, M. J. *et al.* APOBEC-mediated Mutagenesis as a Likely Cause of FGFR3 S249C Mutation Over-representation in Bladder Cancer. *Eur Urol* **76**, 9-13, doi:10.1016/j.eururo.2019.03.032 (2019).
56 Flasch, D. A. *et al.* Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication. *Cell* **177**, 837-851.e828, doi:10.1016/j.cell.2019.02.050 (2019).
57 Sultana, T. *et al.* The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. *Molecular cell* **74**, 555-570.e557, doi:10.1016/j.molcel.2019.02.036 (2019).
58 Ming, X. *et al.* Kinetics and mechanisms of mitotic inheritance of DNA methylation and their roles in aging-associated methylome deterioration. *Cell Res*, doi:10.1038/s41422-020-0359-9 (2020).
59 Reijns, M. A. *et al.* Lagging-strand replication shapes the mutational landscape of the genome. *Nature*, doi:10.1038/nature14183 (2015).
60 Daigaku, Y. *et al.* A global profile of replicative polymerase usage. *Nature structural & molecular biology*, doi:10.1038/nsmb.2962 (2015).
61 Clausen, A. R. *et al.* Tracking replication enzymology in vivo by genome-wide mapping of ribonucleotide incorporation. *Nature structural & molecular biology* **22**, 185-191, doi:10.1038/nsmb.2957 (2015).
62 Koh, K. D., Balachander, S., Hesselberth, J. R. & Storici, F. Ribose-seq: global mapping of ribonucleotides embedded in genomic DNA. *Nat Methods* **12**, 251-257, 253 p following 257, doi:10.1038/nmeth.3259 (2015).
63 Zhou, Z. X., Lujan, S. A., Burkholder, A. B., Garbacz, M. A. & Kunkel, T. A. Roles for DNA polymerase δ in initiating and terminating leading strand DNA replication. *Nat Commun* **10**, 3992, doi:10.1038/s41467-019-11995-z (2019).
64 Koyanagi, E. *et al.* Global landscape of replicative DNA polymerase usage in the human genome. *bioRxiv*, 2021.2011.2014.468503, doi:10.1101/2021.11.14.468503 (2021).
65 Pratto, F. *et al.* Meiotic recombination mirrors patterns of germline replication in mice and humans. *Cell* **184**, 4251-4267.e4220, doi:10.1016/j.cell.2021.06.025 (2021).
66 Sriramachandran, A. M. *et al.* Genome-wide Nucleotide-Resolution Mapping of DNA Replication Patterns, Single-Strand Breaks, and Lesions by GLOE-Seq. *Molecular cell* **78**, 975-985 e977, doi:10.1016/j.molcel.2020.03.027 (2020).

67 Kara, N., Krueger, F., Rugg-Gunn, P. & Houseley, J., doi:10.1101/2020.08.10.243931 (2020).
68 Kit Leng Lui, S. *et al.* Monitoring genome-wide replication fork directionality by Okazaki fragment sequencing in mammalian cells. *Nat Protoc* **16**, 1193-1218, doi:10.1038/s41596-020-00454-5 (2021).
69 Audit, B. *et al.* Multiscale analysis of genome-wide replication timing profiles using a wavelet-based signal-processing algorithm. *Nat Protoc* **8**, 98-110, doi:10.1038/nprot.2012.145 (2013).
70 Muller, C. A. *et al.* Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads. *Nat Methods* **16**, 429-436, doi:10.1038/s41592-019-0394-y (2019).
71 Gansauge, M. T. *et al.* Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Res* **45**, e79, doi:10.1093/nar/gkx033 (2017).
72 Salic, A. & Mitchison, T. J. A chemical method for fast and sensitive detection of DNA synthesis in vivo. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 2415-2420, doi:10.1073/pnas.0712168105 (2008).
73 Burgers, P. M. J. & Kunkel, T. A. Eukaryotic DNA Replication Fork. *Annu Rev Biochem* **86**, 417-438, doi:10.1146/annurev-biochem-061516-044709 (2017).
74 DePamphilis, M. L. *Genome duplication*. (New York : Garland Science/Taylor & Francis Group, 2010).
75 Qu, D. *et al.* 5-Ethynyl-2'-deoxycytidine as a new agent for DNA labeling: detection of proliferating cells. *Analytical biochemistry* **417**, 112-121, doi:10.1016/j.ab.2011.05.037 (2011).
76 Ligasova, A. *et al.* Dr Jekyll and Mr Hyde: a strange case of 5-ethynyl-2'-deoxyuridine and 5-ethynyl-2'-deoxycytidine. *Open Biol* **6**, 150172, doi:10.1098/rsob.150172 (2016).
77 Manska, S., Octaviano, R. & Rossetto, C. C. 5-Ethynyl-2'-deoxycytidine and 5-ethynyl-2'-deoxyuridine are differentially incorporated in cells infected with HSV-1, HCMV, and KSHV viruses. *J Biol Chem* **295**, 5871-5890, doi:10.1074/jbc.RA119.012378 (2020).
78 Green, M. R. & Sambrook, J. *Molecular Cloning : a laboratory manual*. 4. edn, (Cold Spring Harbor Laboratory Press, 2012).
79 Giacca, M., Pelizon, C. & Falaschi, A. Mapping replication origins by quantifying relative abundance of nascent DNA strands using competitive polymerase chain reaction. *Methods* **13**, 301-312, doi:Doi 10.1006/Meth.1997.0529 (1997).
80 Tornoe, C. W., Christensen, C. & Meldal, M. Peptidotriazoles on solid phase: [1,2,3]-triazoles by regiospecific copper(i)-catalyzed 1,3-dipolar cycloadditions of terminal alkynes to azides. *J Org Chem* **67**, 3057-3064, doi:10.1021/jo011148j (2002).
81 Rostovtsev, V. V., Green, L. G., Fokin, V. V. & Sharpless, K. B. A Stepwise Huisgen Cycloaddition Process: Copper(I)-Catalyzed Regioselective "Ligation" of Azides and Terminal Alkynes. *Angewandte Chemie International Edition* **41**, 2596-2599, doi:10.1002/1521-3773(20020715)41:14<2596::Aid-anie2596>3.0.Co;2-4 (2002).
82 Presolski, S. I., Hong, V. P. & Finn, M. G. Copper-Catalyzed Azide-Alkyne Click Chemistry for Bioconjugation. *Curr Protoc Chem Biol* **3**, 153-162, doi:10.1002/9780470559277.ch110148 (2011).
83 Kwok, C. K., Ding, Y., Sherlock, M. E., Assmann, S. M. & Bevilacqua, P. C. A hybridization-based approach for quantitative and low-bias single-stranded DNA ligation. *Analytical biochemistry* **435**, 181-186, doi:10.1016/j.ab.2013.01.008 (2013).

84 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226, doi:10.1126/science.1224344 (2012).
85 Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165, doi:10.1093/nar/gkw257 (2016).
86 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).
87 Team, R. C. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2021 (2021).
88 Himmelmann, L. HMM: HMM-Hidden Markov Models. R package version 1.0 (2016).
89 Morgan, M., Pages, H., Obenchain, V. & Hayden, N. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. R package version 1.30. 0 (2017).
90 Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS computational biology* **9**, e1003118, doi:10.1371/journal.pcbi.1003118 (2013).
91 Team, R. S. RStudio: Integrated Development for Rstudio Team (PBC Boston, MA, 2020).
92 Ma, E., Hyrien, O. & Goldar, A. Do replication forks control late origin firing in Saccharomyces cerevisiae? *Nucleic acids research* **40**, 2010-2019, doi:10.1093/nar/gkr982 (2011).

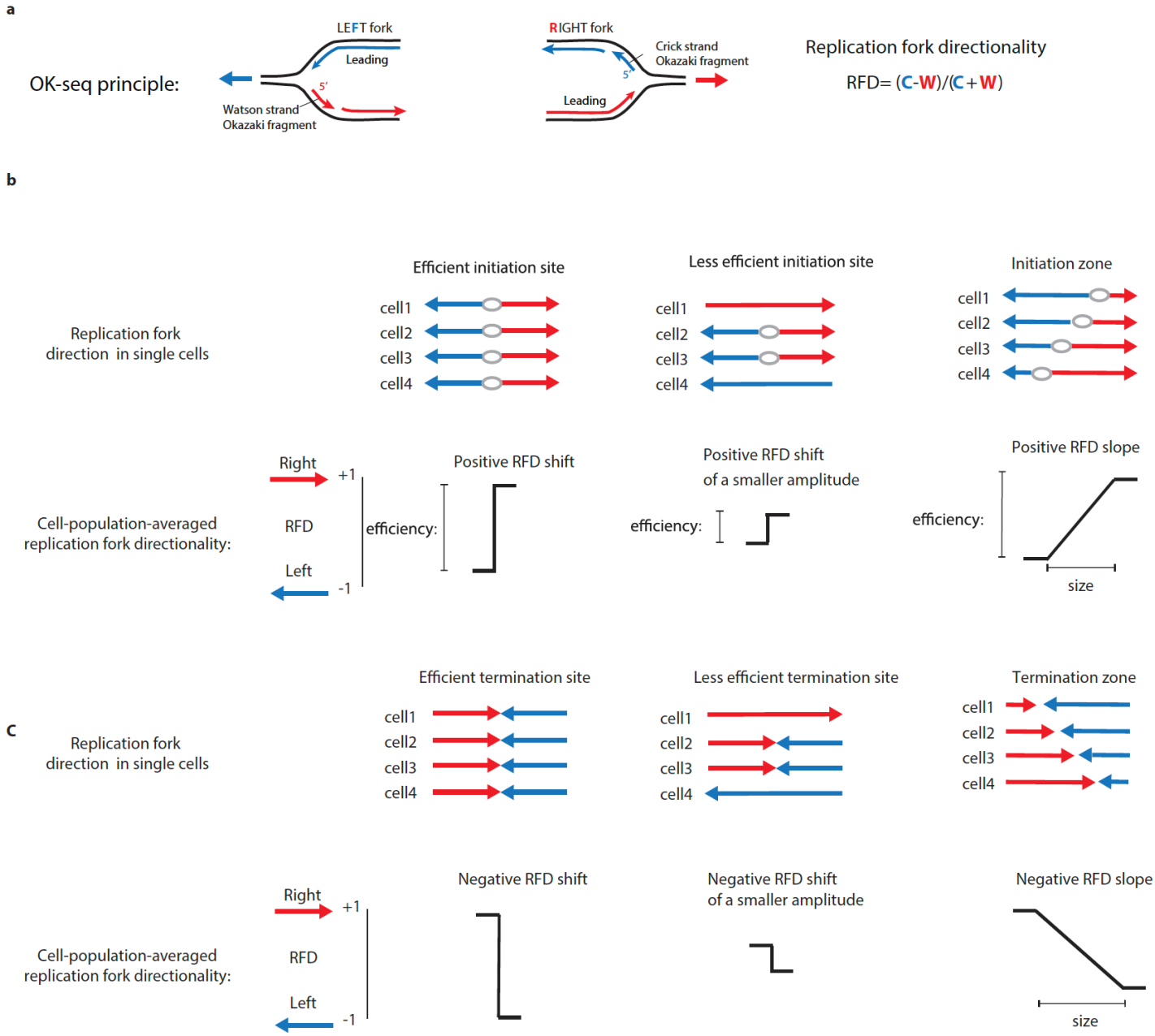
a
OK-seq principle:
LEFT fork
Leading
5'
Watson strand
Okazaki fragment
RIGHT fork
Crick strand
Okazaki fragment
5'
Leading
Replication fork directionality
RFD= (C-W)/(C+W)
b
Replication fork
direction in single cells
Efficient initiation site
Less efficient initiation site
Initiation zone
cell1
cell2
cell3
cell4
Cell-population-averaged
replication fork directionality:
Right
+1
RFD
Left
-1
efficiency:
Positive RFD shift
Positive RFD shift
of a smaller amplitude
Positive RFD slope
size
c
Replication fork
direction in single cells
Efficient termination site
Less efficient termination site
Termination zone
Cell-population-averaged
replication fork directionality:
Negative RFD shift
Negative RFD shift
of a smaller amplitude
Negative RFD slope
size

Figure 1

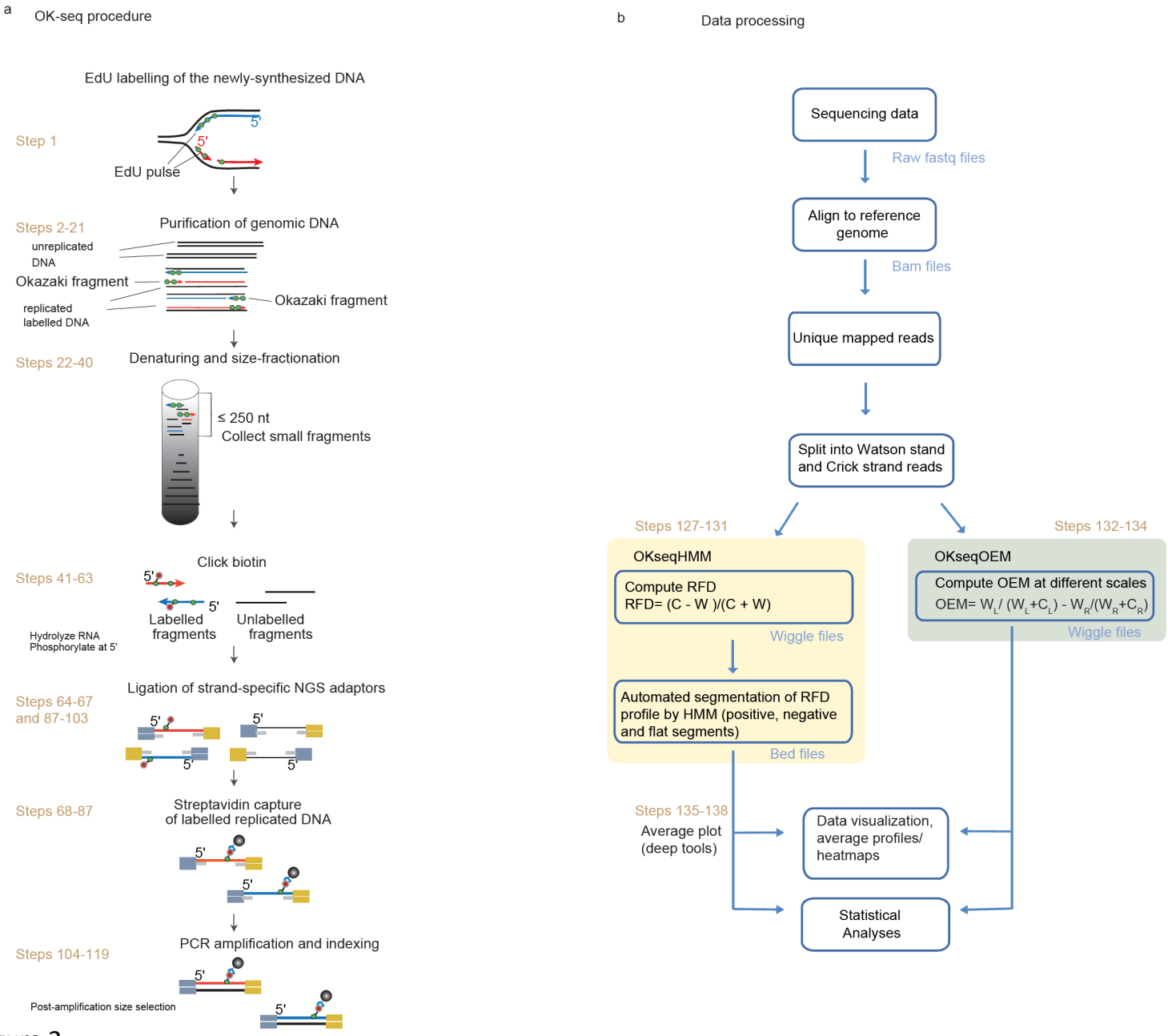
a
OK-seq procedure
EdU labelling of the newly-synthesized DNA
Step 1
5'
5'
EdU pulse
Steps 2-21
Purification of genomic DNA
unreplicated DNA
Okazaki fragment
Okazaki fragment
replicated labelled DNA
Steps 22-40
Denaturing and size-fractionation
≤ 250 nt
Collect small fragments
Click biotin
Steps 41-63
5'
5'
Labelled fragments
Unlabelled fragments
Hydrolyze RNA
Phosphorylate at 5'
Ligation of strand-specific NGS adaptors
Steps 64-67 and 87-103
5'
5'
5'
5'
Steps 68-87
Streptavidin capture of labelled replicated DNA
5'
5'
PCR amplification and indexing
Steps 104-119
5'
5'
Post-amplification size selection
b
Data processing
Sequencing data
Raw fastq files
Align to reference genome
Bam files
Unique mapped reads
Split into Watson stand and Crick strand reads
Steps 127-131
OKseqHMM
Compute RFD
RFD= (C - W)/(C + W)
Wiggle files
Automated segmentation of RFD profile by HMM (positive, negative and flat segments)
Bed files
Steps 132-134
OKseqOEM
Compute OEM at different scales
OEM= W_L/ (W_L+C_L) - W_R/(W_R+C_R)
Wiggle files
Steps 135-138
Average plot (deep tools)
Data visualization, average profiles/ heatmaps
Statistical Analyses

Figure 2

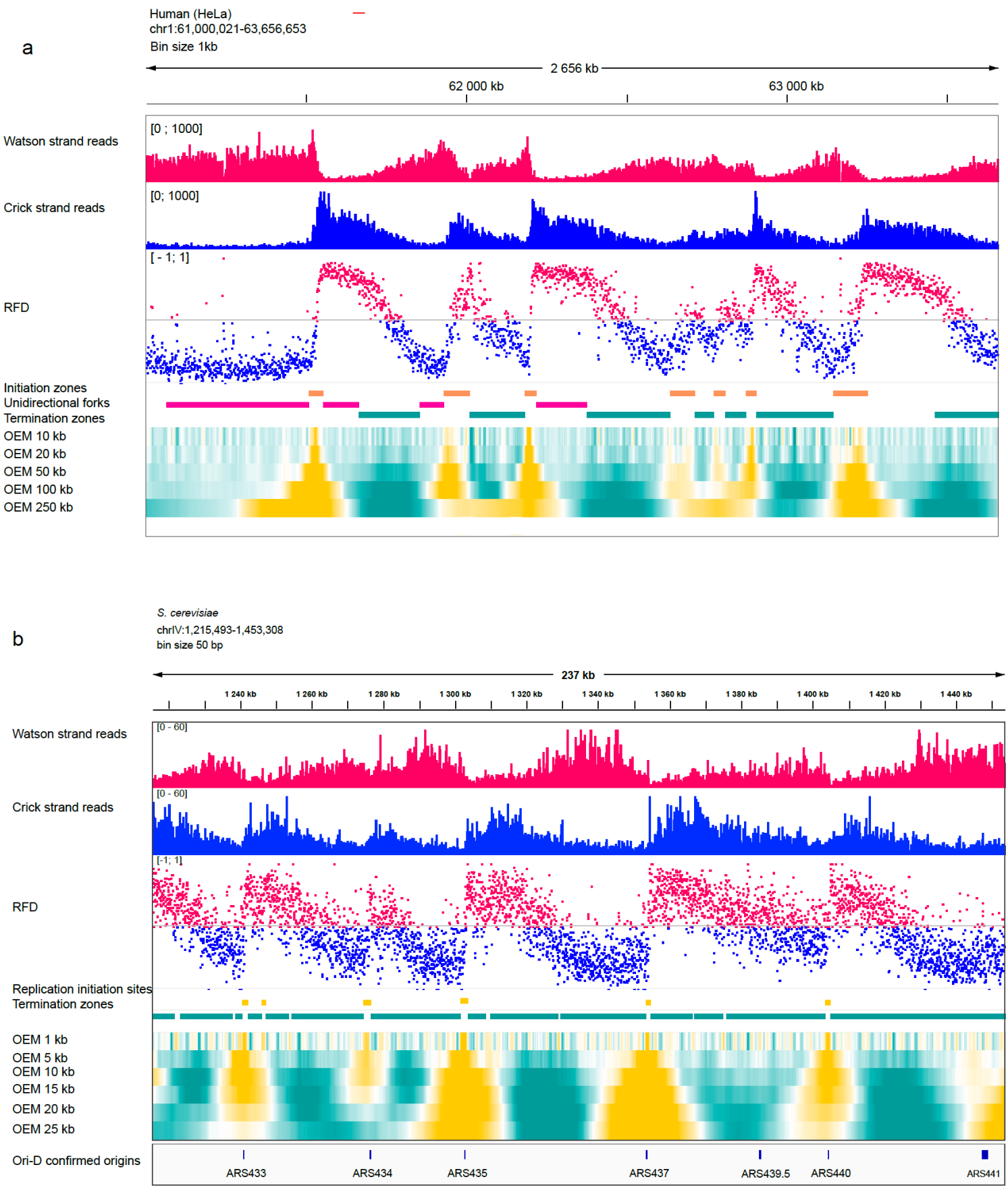
a
Human (HeLa)
chr1:61,000,021-63,656,653
Bin size 1kb
2 656 kb
62 000 kb
63 000 kb
Watson strand reads
[0 ; 1000]
Crick strand reads
[0; 1000]
[- 1; 1]
RFD
Initiation zones
Unidirectional forks
Termination zones
OEM 10 kb
OEM 20 kb
OEM 50 kb
OEM 100 kb
OEM 250 kb
b
S. cerevisiae
chrIV:1,215,493-1,453,308
bin size 50 bp
237 kb
1 240 kb
1 260 kb
1 280 kb
1 300 kb
1 320 kb
1 340 kb
1 360 kb
1 380 kb
1 400 kb
1 420 kb
1 440 kb
Watson strand reads
[0 - 60]
Crick strand reads
[0 - 60]
[-1; 1]
RFD
Replication initiation sites
Termination zones
OEM 1 kb
OEM 5 kb
OEM 10 kb
OEM 15 kb
OEM 20 kb
OEM 25 kb
Ori-D confirmed origins
ARS433
ARS434
ARS435
ARS437
ARS439.5
ARS440
ARS441

Figure 3

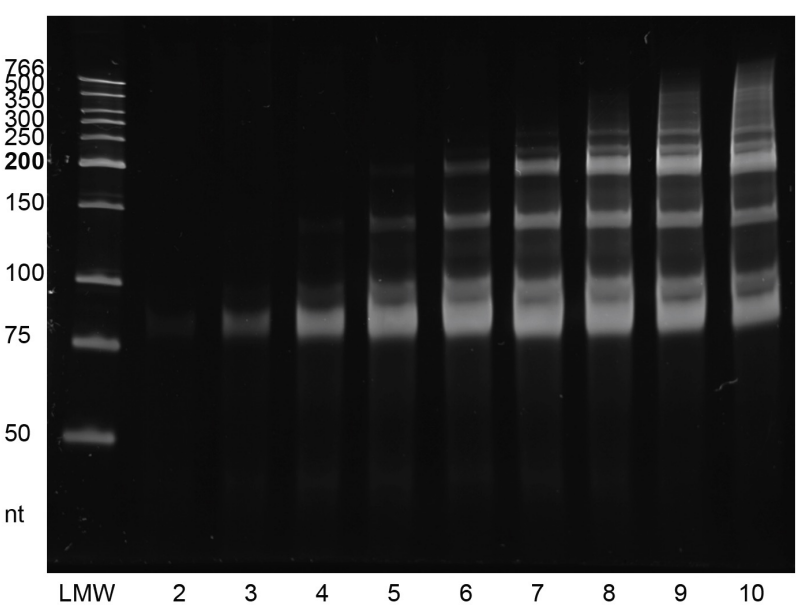

Figure Box 1

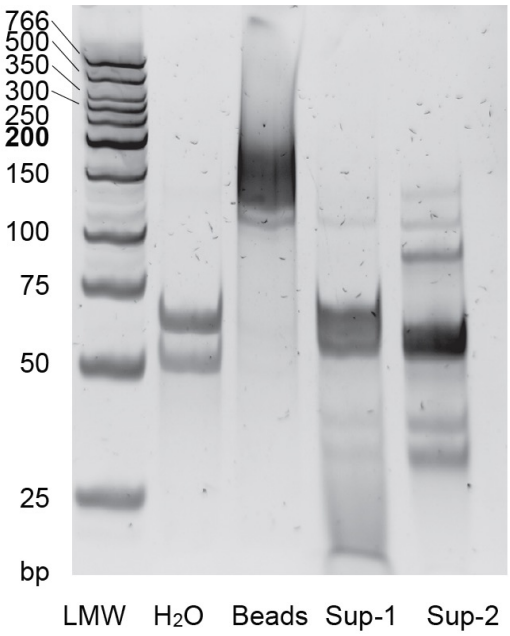

Figure Box 3

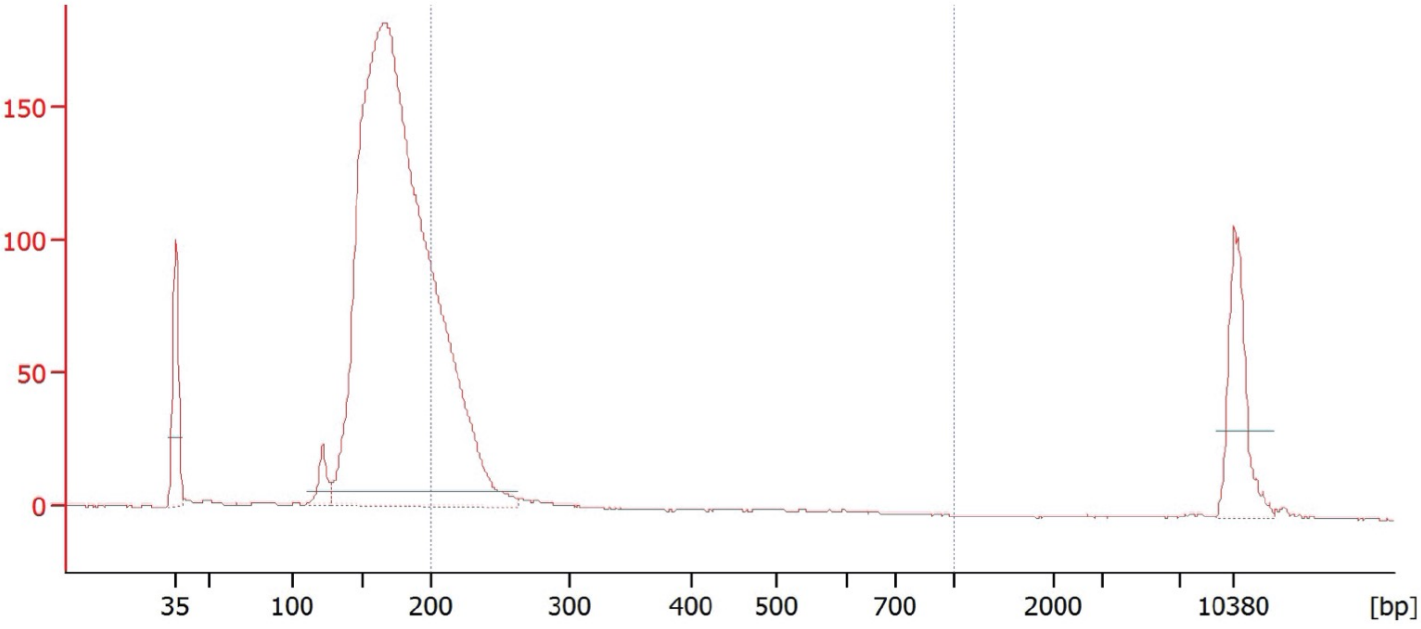

Figure Box 4

Annex 2

Topoisomerase I prevents transcription-replication conflicts at transcription termination sites

Yaqun Liu , Yea-Lih Lin , Philippe Pasero & Chun-Long Chen

(Published in *Molecular & Cellular Oncology*)

Annex 3

The impact of transcription-mediated replication stress on genome instability and human disease

Stefano Gnan, Yaqun Liu, Manuela Spagnuolo, Chun-Long Chen

(Published in *Genome instability & Disease*)

Annex 4

The corresponding scripts in Chapter 2 & 3

Scripts in Chapter 2:

OKseqHMM.R

```
#' OKseqHMM backage
#' This function allows you to generate the two corresponding strand bam files, to generate the RFD profiles and to identify most of the replication initiation/termination zones and also the intermediate states which RFD profiles are noramelly flat.
#' @keywords OK-Seq, RFD, peak calling, HMM
#' @export
#' @examples
#' OKseqHMM()

#       Initialize HMM 4 states, observations, start probability, emission probability, transition probability

OKseqHMM <- function(bamfile,chrsizes,fileOut, thresh, winS, binSize, hwinS=winS/2,
                     st=c("D", "L", "H", "U"),
                     sym=c("V", "W", "X", "Y", "Z"),
                     pstart=rep(1/4, 4),

pem=t(matrix(c(0.383886256,0.255924171,0.170616114,0.113744076,0.075829384,
                               .10,.20,.40,.20,.10,
                               .10,.20,.40,.20,.10,
                               0.022222222,      0.033333333,      0.066666667,      0.211111111,
0.666666667),
                             ncol=4)),
                     ptrans=t(matrix(c(0.9999,0.000020,0,0.000080,
                                       0,0.999,0,0.001,
                                       0.001,0,0.999,0,
                                       0.000080,0,0.000020,0.9999),
                                     ncol=4)),
                     quant=c(-1,              -0.0082058939609862,              -0.00141890249101162,
0.00103088286465956, 0.00800467305420799, 1))

{
```

```
require(HMM)
require(Rsamtools)
require(GenomicAlignments)

readBAM<- readGAlignments(bamfile)
chrom <- as.character(na.omit(unique(seqnames(readBAM))))
print(chrom)
paired <- testPairedEndBam(bamfile)
# test if the BAM file is pair-end or not

if (paired)
{
  #Generate forward and reverse strand bam files:
  print("This bam is pair-end.")
  print("Seperating the forward strand bam.")
  # include reads that are 2nd in a pair (128);
  # exclude reads that are mapped to the reverse strand (16)
  system(paste0("samtools view -b -f 128 -F 16 ",bamfile," > a.fwd1.bam"))

  # include reads that are mapped to the reverse strand (16) and
  # first in a pair (64): 64 + 16 = 80
  system(paste0("samtools view -b -f 80 ",bamfile," > a.fwd2.bam"))

  # combine the temporary files
  system(paste0("samtools merge -f ",fileOut,"_fwd.bam a.fwd1.bam a.fwd2.bam"))
  system(paste0("samtools index ",fileOut,"_fwd.bam"))

  # remove the temporary files
  system(paste0("rm a.fwd*.bam"))

  print("Seperating the reverse strand bam.")
  # include reads that map to the reverse strand (128)
  # and are second in a pair (16): 128 + 16 = 144
  system(paste0("samtools view -b -f 144 ",bamfile," > a.rev1.bam"))

  # include reads that are first in a pair (64), but
  # exclude those ones that map to the reverse strand (16)
```

```
  system(paste0("samtools view -b -f 64 -F 16 ",bamfile," > a.rev2.bam"))

  # merge the temporary files
  system(paste0("samtools merge -f ",fileOut,"_rev.bam a.rev1.bam a.rev2.bam"))

  # index the merged, filtered BAM file
  system(paste0("samtools index ",fileOut,"_rev.bam"))
  # remove temporary files
  system(paste0("rm a.rev*.bam"))
 }
 else
 {
  print("This bam is single-end.")

  print("Seperating the forward strand bam.")
  # Forward strand.
  system(paste0("samtools view -bh -f 16 ",bamfile," > ",fileOut,"_fwd.bam"))
  system(paste0("samtools index ",fileOut,"_fwd.bam"))

  print("Seperating the reverse strand bam.")
  # Reverse strand
  system(paste0("samtools view -bh -F 16 ",bamfile," > ",fileOut,"_rev.bam"))
  system(paste0("samtools index ",fileOut,"_rev.bam"))

 }

 chromNames <- read.table(chrsizes,header=FALSE,sep="\t",comment.char =
"#",stringsAsFactors = FALSE)
 chr.sizes <- data.frame(chr=chromNames[,1],size=chromNames[,2])

 for (i in c(1:length(chrom))){
  chr.name <- chrom[i]
  print(chr.name)
  chr.length <- chr.sizes[chr.sizes$chr == chr.name,2]
  print(chr.length)
  print(paste0("Calculating ",binSize/1000,"kb binsize coverage for forward strand."))
```

```
    system(paste0("samtools view ",fileOut,"_fwd.bam ",chr.name," > fwd_",chr.name,".sam"))
    system(paste0("awk '$3~/^", chr.name, "$/ {print $2 \"\t\" $4}' fwd_",chr.name,".sam > fwd_",chr.name,".txt"))
    fileIn <- paste0("fwd_",chr.name,".txt")
    tmp <- read.table(fileIn, header=F, comment.char="",colClasses=c("integer","integer"),fill=TRUE)
    tags <- tmp[,2]
    tags[tags<=0] <- 1
    breaks <- seq(0, chr.length+binSize, by=binSize)
    h <- hist(tags, breaks=breaks, plot=FALSE)
    c <- h$counts

    print(paste0("Calculating ",binSize/1000,"kb binsize coverage for reverse strand."))
    system(paste0("samtools view ",fileOut,"_rev.bam ",chr.name," > rev_",chr.name,".sam"))
    system(paste0("awk '$3~/^", chr.name, "$/ {print $2 \"\t\" $4}' rev_",chr.name,".sam > rev_",chr.name,".txt"))
    fileIn <- paste0("rev_",chr.name,".txt")
    tmp<-
read.table(fileIn,header=F,comment.char="",colClasses=c("integer","integer"),fill=TRUE)
    tags <- tmp[,2]
    tags[tags<=0] <- 1
    breaks <- seq(0, chr.length+binSize, by=binSize)
    h <- hist(tags, breaks=breaks, plot=FALSE)
    w <- h$counts
    system(paste0("rm *.sam"))
    system(paste0("rm f*.txt"))
    system(paste0("rm r*.txt"))

    # raw polarity for later
    polar <- c/(c+w)
    polar[c<thresh & w<thresh] <- NA

    # 1kb RFD:
    rfd <- (c-w)/(w+c)
    rfd[is.na(rfd)] <- 0
    rfd[w<thresh & c<thresh] <- 0
    rfd[rfd > 1] <- 1
```

```
    rfd[rfd < -1] <- -1

    start_pos <- as.integer(breaks[1:length(breaks)-1])
    end_pos <- as.integer(breaks[2 : length(breaks)])
    chrName <- rep(chr.name,length(rfd))
    df <-data.frame(chr=chrName,startPos = start_pos,endPos = end_pos,rd_nb=rfd)
    #restrict the last position is the chr.length, not exceed that.
    df$endPos[nrow(df)] <- chr.length
    write.table(df, file = paste0(fileOut,"_RFD_cutoff",thresh,"_bs",binSize/1000,"kb.bedgraph",
sep=""), append = T, quote = FALSE, sep = "\t", col.names=F, row.names=F)

    # smoothing RFD
    print(paste("Smoothing window size :", winS, "kb"))
    sw <- cumsum(w)
    lg <- length(w)
    from <- (-hwinS+2):(lg-hwinS+1)
    to <- from+winS-1
    from[from<1] <- 1
    to[to>lg] <- lg

    print("")
    print(paste("number of bins :", length(w)))
    print("")

    win <- matrix(c(from, to), ncol=2)
    ws <- apply(win ,1, function(x) { (sw[x[2]]-sw[x[1]])/winS } )

    sc <- cumsum(c)
    lg <- length(c)
    cs <- apply(win ,1, function(x) { (sc[x[2]]-sc[x[1]])/winS } )

    print(paste("cutoff is :",thresh))
    rfd <- (cs-ws)/(ws+cs)
    rfd[is.na(rfd)] <- 0
    rfd[ws<thresh & cs<thresh] <- 0
    rfd[rfd > 1] <- 1
    rfd[rfd < -1] <- -1
```

```
    start_pos <- as.integer(breaks[1:length(breaks)-1])
    end_pos <- as.integer(breaks[2 : length(breaks)])
    chrName <- rep(chr.name,length(rfd))
    df <-data.frame(chr=chrName,startPos = start_pos,endPos = end_pos,rd_nb=rfd)
    #restrict the last position is the chr.length, not exceed that.
    df$endPos[nrow(df)] <- chr.length
    write.table(df,                                    file                                    =
paste0(fileOut,"_RFD_cutoff",thresh,"_bs",binSize/1000,"kb_sm_",winS,"kb.bedgraph",
sep=""), append = T, quote = FALSE, sep = "\t", col.names=F, row.names=F)

    # HMM from new deltas ============================

    # derive
    bias <- cs/(ws+cs)
    bias[is.na(bias)] <- 0.5
    bias[ws<thresh & cs<thresh] <- 0.5

    delta <- c(0,bias[-1]-bias[-length(bias)])
    delta[is.na(delta)] <- 0.5

    # affect symbols
    if (is.na(quant[1])) { quant <- quantile(delta, probs = seq(0, 1, 0.20)) }
    quant[1] <- -1
    quant[length(quant)] <- 1

    print("quantile borders :")
    print(quant)
    print("")
    dx  <- unlist(sapply(delta, function(x) { ix <- which(x>=quant); ix[length(ix)] }))
    dx[dx>5] <- 5

    # write log ================================

    logFile <- paste(fileOut,"_log.txt", sep="")

    write.table(data.frame(c("fileOut",fileOut)), file = logFile,
```

```
        append = T, quote = FALSE, sep = "\t", col.names=F, row.names=F)
    write.table(paste("ptrans",chr.name), file = logFile,
        append = T, quote = FALSE, sep = "\t", col.names=F, row.names=F)
    write.table(ptrans, file = logFile,
        append = T, quote = FALSE, sep = "\t", col.names=F, row.names=F)
    write.table(paste("pem",chr.name), file = logFile,
        append = T, quote = FALSE, sep = "\t", col.names=F, row.names=F)
    write.table(pem, file = logFile,
        append = T, quote = FALSE, sep = "\t", col.names=F, row.names=F)
    write.table(data.frame("pstart",chr.name), file = logFile,
        append = T, quote = FALSE, sep = "\t", col.names=F, row.names=F)
    write.table(t(data.frame(pstart)), file = logFile,
        append = T, quote = FALSE, sep = "\t", col.names=F, row.names=F)
    write.table(data.frame("st",chr.name), file = logFile,
        append = T, quote = FALSE, sep = "\t", col.names=F, row.names=F)
    write.table(t(data.frame(st)), file = logFile,
        append = T, quote = FALSE, sep = "\t", col.names=F, row.names=F)
    write.table(data.frame("sym",chr.name), file = logFile,
        append = T, quote = FALSE, sep = "\t", col.names=F, row.names=F)
    write.table(t(data.frame(sym)), file = logFile,
        append = T, quote = FALSE, sep = "\t", col.names=F, row.names=F)
    write.table(data.frame("quant"), file = logFile,
        append = T, quote = FALSE, sep = "\t", col.names=F, row.names=F)
    write.table(quant, file = logFile,
        append = T, quote = FALSE, sep = "\t", col.names=F, row.names=F)

    # HMM
    hmm1 <- initHMM(States=st, Symbols=sym, startProbs=pstart, transProbs=ptrans,
emissionProbs=pem)
    print(hmm1)
    print("wait, viterbi...")
    seg <- viterbi(hmm=hmm1, observation=dx)

    # pour affichage du profil des etats
    prof <- rep(NA, times=length(seg))
    prof[seg=="U"] <- 1
    prof[seg=="H"] <- 0.2
```

```
    prof[seg=="L"] <- -0.2
    prof[seg=="D"] <- -1
    prof[is.na(w[-length(w)])] <- -1.1
    prof[is.na(c[-length(c)])] <- -1.1
    prof[is.na(prof)] <- -.1

    write.table(prof, file = paste(fileOut,"_HMM.txt", sep=""), append = T, quote = FALSE, sep =
"\t", col.names=F, row.names=F)

    # adding the probability curve
    print("wait, probabilities…")
    post <- posterior(hmm1,dx)
    prb <- rep(NA, length(seg))
    for (i in 1:length(seg))
    {
      prb[i] <- post[seg[i],i]
    }
    prb[prb>1] <- 1

    write.table(as.integer(prb*1000), file = paste(fileOut, "_HMMproba.txt", sep=""), append = T,
                quote = FALSE, sep = "\t", col.names=F, row.names=F)

    # calculate the region coordinates by profil viterbi ========================

    left <- seg[-length(seg)]
    right <- seg[-1]
    ix <- which(left!=right)
    from <- c(1,ix+1)
    to <- c(ix,length(seg))

    # for the table, adding states and calculating lengths, slopes and associated probabilities
    # calculate the slope of the polarity of the states (between the left and right part of the state))

    states <- meanPol <- ymax <- ymin <- p <- inc <- napc <- corr <- rep(NA, length(from))
    states <- seg[from]

    for (i in 1:length(from))
```

```
{
  pos <- from[i]:to[i]
  lgpos <- length(pos)

  m <- prb[pos[!is.na(prb[pos])]]
  if (length(m)>0)
  {
    p[i] <- mean(m, na.rm=T)
  }
  meanPol[i] <- mean(polar[pos], na.rm=T)

  pos2 <- pos[!is.na(polar[pos])]
  lgpos2 <- length(pos2)
  napc[i] <- 100-round(lgpos2/lgpos*100)
  if (lgpos2<2)
  {
    res <- data.frame(NA, NA)
    inc[i] <- ymin[i] <- ymax[i] <- NA
  } else {
    realPos <- pos2*binSize
    res <- lm(polar[pos2] ~ realPos)
    inc[i] <- res[[1]][2]
    ymin[i] <-          (inc[i]*realPos[1])+res[[1]][1]
    ymax[i] <-          (inc[i]*realPos[lgpos2])+res[[1]][1]
    corr[i] <- cor(x=realPos, y=polar[pos2])
  }
}
ymin[ymin<0] <- 0
ymax[ymax<0] <- 0
ymin[ymin>1] <- 1
ymax[ymax>1] <- 1

# for display

inc1 <- round(inc*10^8)              # as a percentage of RFD per megabase
p <- round(p*100)
chr <- rep(chr.name,length(from))
```

```
    from1 <- (from-1)*binSize+1
    to1 <- to*binSize
    lg1 <- to1-from1+1
    meanPol1 <- round(meanPol*100)
    ymin1 <- round(ymin*100)
    ymax1 <- round(ymax*100)
    corr1 <- round(corr*100)

    # adjustment of the border polarities by the average and recalculation of the slope

    polL <- ymin1
    polR <- ymax1

    polD <- polR[-length(polR)]
    polG <- polL[-1]

    polD[is.na(polD)] <- polG[is.na(polD)]
    polG[is.na(polG)] <- polD[is.na(polG)]

    polM <- round((polG+polD)/2)
    polL[-1] <- polR[-length(polR)] <- polM

    # adjusted slope: 10 ^ 6 for clarity of the table display (% by megabase)
    slope_adj <- round(10^6*(polR-polL)/(to1-from1))

    # writing
    dataOut <- data.frame(chr, from=as.integer(from1), to=as.integer(to1), state=states,
length=lg1, slope=inc1,
                          p, pol_mean=meanPol1, pol_left=ymin1, pol_right=ymax1,
                          na=napc, cor=corr1, slope_adj=slope_adj, pol_adj_left=polL,
pol_adj_right=polR)
    dataOut_U <- dataOut[dataOut$state == "U",]
    dataOut_D <- dataOut[dataOut$state == "D",]
    dataOut_HF <- dataOut[dataOut$state == "H",]
    dataOut_LF <- dataOut[dataOut$state == "L",]
```

```
   write.table(dataOut_U, file = paste(fileOut,"_HMMsegments_IZ.txt", sep=""), append = T,
               quote = FALSE, sep = "\t", col.names=T, row.names=F)
   write.table(dataOut_U[,1:3], file = paste(fileOut,"_HMMsegments_IZ.bed", sep=""), append
= T,
               quote = FALSE, sep = "\t", col.names=F, row.names=F)

   write.table(dataOut_D, file = paste(fileOut,"_HMMsegments_TZ.txt", sep=""), append = T,
               quote = FALSE, sep = "\t", col.names=T, row.names=F)
   write.table(dataOut_D[,1:3], file = paste(fileOut,"_HMMsegments_TZ.bed", sep=""), append
= T,
               quote = FALSE, sep = "\t", col.names=F, row.names=F)

   write.table(dataOut_HF, file = paste(fileOut,"_HMMsegments_highFlatZone.txt", sep=""),
append = T,
               quote = FALSE, sep = "\t", col.names=T, row.names=F)
   write.table(dataOut_HF[,1:3], file = paste(fileOut,"_HMMsegments_highFlatZone.bed",
sep=""), append = T,
               quote = FALSE, sep = "\t", col.names=F, row.names=F)

   write.table(dataOut_LF, file = paste(fileOut,"_HMMsegments_LowFlatZone.txt", sep=""),
append = T,
               quote = FALSE, sep = "\t", col.names=T, row.names=F)
   write.table(dataOut_LF[,1:3], file = paste(fileOut,"_HMMsegments_LowFlatZone.bed",
sep=""), append = T,
               quote = FALSE, sep = "\t", col.names=F, row.names=F)
  }

}
# end of the function
```

OKseqOEM.R

```
OKseqOEM <- function(bamInF, bamInR, chrsizes, fileOut, binSize, binList)
{
 chromNames <- read.table(chrsizes,header=FALSE,sep="\t",comment.char = "#",stringsAsFactors = FALSE)
 chr.sizes <- data.frame(chr=chromNames[,1],size=chromNames[,2])
 require(Rsamtools)
 paired <- testPairedEndBam(bamInF)
 for (i in c(1:nrow(chr.sizes))){
  chr.name <- chr.sizes$chr[i]
  print(chr.name)
  chr.length <- chr.sizes$size[i]
  print(chr.length)
  if (paired)
  {
   print(paste0("It's pair-end. Calculating ",binSize,"bp binsize coverage for forward strand."))
   system(paste0("samtools view -q 1 -f 0x42 -F 0x4 ",bamInF," ",chr.name," > fwd_",chr.name,".sam"))
   system(paste0("awk '$3~/^", chr.name, "$/ {print $2 \"\t\" $4}' fwd_",chr.name,".sam > fwd_",chr.name,".txt"))

   print(paste0("Calculating ",binSize,"bp binsize coverage for reverse strand."))
   system(paste0("samtools view -q 1 -f 0x42 -F 0x4 ",bamInR," ",chr.name," > rev_",chr.name,".sam"))
   system(paste0("awk '$3~/^", chr.name, "$/ {print $2 \"\t\" $4}' rev_",chr.name,".sam > rev_",chr.name,".txt"))
  }else{
   print(paste0("It's single-end. Calculating ",binSize,"bp binsize coverage for forward strand."))
   system(paste0("samtools view ",bamInF," ",chr.name," > fwd_",chr.name,".sam"))
   system(paste0("awk '$3~/^", chr.name, "$/ {print $2 \"\t\" $4}' fwd_",chr.name,".sam > fwd_",chr.name,".txt"))

   print(paste0("Calculating ",binSize,"bp binsize coverage for reverse strand."))
   system(paste0("samtools view ",bamInR," ",chr.name," > rev_",chr.name,".sam"))
```

```
    system(paste0("awk '$3~/^", chr.name, "$/ {print $2 \"\t\" $4}' rev_",chr.name,".sam > rev_",chr.name,".txt"))
  }
  fileIn <- paste0("fwd_",chr.name,".txt")
  tmp<- read.table(fileIn, header=F, comment.char="",colClasses=c("integer","integer"),fill=TRUE)
  tags <- tmp[,2]
  tags[tags<3] <- 0
  breaks <- seq(0, chr.length+binSize, by=binSize)
  h <- hist(tags, breaks=breaks, plot=FALSE)
  Temp.chr.F <- h$counts
  fileIn <- paste0("rev_",chr.name,".txt")
  tmp <- read.table(fileIn, header=F, comment.char="",colClasses=c("integer","integer"),fill=TRUE)
  tags <- tmp[,2]
  tags[tags<3] <- 0
  breaks <- seq(0, chr.length+binSize, by=binSize)
  h <- hist(tags, breaks=breaks, plot=FALSE)
  Temp.chr.R <- h$counts

  system(paste0("rm *.sam"))
  system(paste0("rm f*.txt"))
  system(paste0("rm r*.txt"))

  Temp.chr.F <- cumsum(Temp.chr.F)
  Temp.chr.R <- cumsum(Temp.chr.R)

  print("Calculating OEM.")
  for (n in c(1:length(binList)))
  {

    print(paste0("The smoothing window size for OEM is ",binList[n]*binSize/1000,"kb."))

    Data.chr.F <- Temp.chr.F[(binList[n]+1):length(Temp.chr.F)]-Temp.chr.F[1:(length(Temp.chr.F)-binList[n])]
    Data.chr.R <- Temp.chr.R[(binList[n]+1):length(Temp.chr.R)]-Temp.chr.R[1:(length(Temp.chr.R)-binList[n])]
```

```
    Data.chr.Smooth <- Data.chr.F/(Data.chr.F+Data.chr.R)
    Data.chr <- Data.chr.Smooth[(binList[n]+1):length(Data.chr.Smooth)]-
Data.chr.Smooth[1:(length(Data.chr.Smooth)-binList[n])]

    Data.chr <- c(rep(NA,binList[n]-1),Data.chr)
    Data.chr[which(is.na(Data.chr))] <- 0
    # Data.chr<- Data.chr[1:chr.length]

    ##Save file in wig format
    if (i==1) {
      Title <- paste0("fixedStep chrom=", chr.name, " start=1 step=",binSize," span=",binSize,
sep="")
      fileOutWig <- paste0(fileOut,"_OEM_",binList[n]*binSize/1000,"kb.wig")
      write.table(Data.chr, file=fileOutWig, quote = FALSE, row.names = FALSE,
col.names=Title, append = FALSE)
    } else {

      Title <- paste0("fixedStep chrom=", chr.name, " start=1 step=",binSize," span=",binSize,
sep="")
      fileOutWig <- paste0(fileOut,"_OEM_",binList[n]*binSize/1000,"kb.wig")
      write.table(Data.chr, file=fileOutWig, quote = FALSE, row.names = FALSE,
col.names=Title, append = TRUE)

    }
   }
  }
}
```

Average_profile_heatmap.sh

PLOT average profile of RFD around TSS and TTS

computeMatrix scale-regions --regionsFileName {your bed file of interested regions/genes PATH e.g.codingGenes.bed} --beforeRegionStartLength {e.g. 10000} --afterRegionStartLength {e.g. 10000} --regionBodyLength {e.g. 20000} --binSize {e.g. 1000} --scoreFileName {RFD bigwig file PATH e.g. Hela.EdC.Combined_OkaSeq.RFD.bw} --outFileName {e.g. "OUTPUT.matrix"} --missingDataAsZero --skipZeros

plotProfile --matrixFile {e.g. "OUTPUT.matrix"} --outFileName {e.g. "RFD_averageProfile.stGeneLength.png"} --averageType mean --startLabel {e.g. start/TSS} --endLabel {e.g. end/TTS} --plotType se
##
#

PLOT Heatmap of RFD IZ center +/-100kb
computeMatrix reference-point --regionsFileName {your IZ bed file PATH e.g. HeLa_hmm_HMMsegments_IZ.bed} --beforeRegionStartLength {e.g. 100000} --afterRegionStartLength {e.g. 100000} --binSize {e.g. 1000} --scoreFileName {RFD bigwig file PATH e.g. Hela.EdC.Combined_OkaSeq.RFD.bw} --outFileName {e.g. "OUTPUT.matrix"} --missingDataAsZero --skipZeros --referencePoint center
plotHeatmap --matrixFile {e.g. "OUTPUT.matrix"} --outFileName {e.g. "RFD_sortbyLength.png"} --whatToShow "plot, heatmap and colorbar" --samplesLabel {e.g. HeLa} --refPointLabel center --sortUsing region_length --sortRegions ascend

PLOT Heatmap of OEM IZ center +/-100kb
computeMatrix reference-point --regionsFileName {your IZ bed file PATH e.g. HeLa_hmm_HMMsegments_IZ.bed} --beforeRegionStartLength {e.g. 100000} --afterRegionStartLength {e.g. 100000} --binSize {e.g. 1000} --scoreFileName {series of OEM bigwig file PATH e.g. 20130819CGM130726.Hela_OEM_10kb.bw 20130819CGM130726.Hela_OEM_20kb.bw 20130819CGM130726.Hela_OEM_50kb.bw 20130819CGM130726.Hela_OEM_100kb.bw 20130819CGM130726.Hela_OEM_250kb.bw 20130819CGM130726.Hela_OEM_500kb.bw 20130819CGM130726.Hela_OEM_1Mb.bw} --outFileName {e.g. "OUTPUT.matrix"} --missingDataAsZero --skipZeros --referencePoint center

plotHeatmap --matrixFile {e.g. "OUTPUT.matrix"} --outFileName {e.g. "OEM_sortbyLength.png"} --whatToShow "plot, heatmap and colorbar" --refPointLabel center --samplesLabel {e.g. "HeLa 10kb" "HeLa 20kb" "HeLa 50kb" "HeLa 100kb" "HeLa 250kb" "HeLa 500kb" "HeLa 1Mb"} --sortUsing region_length --sortRegions ascend

Scripts for Chapter 3

#################### Fig.1f - all coding genes with error band ####################

in R:

```
rf_Data <- read.csv('01_count_RPKM_and_peaks_intersect_on_refSeq_hg19.csv', sep = '\t', header = TRUE)
colnames(rf_Data)[4] <- "mRNA"
hela_rnaRPKM <- rf_Data[,c(1,2,3,4,6,60,63)]
mRNA2name <- read.table("mRNA2geneName.txt",header = FALSE,col.names = c("mRNA","chr","strand","geneName"))
mRNA2name <- na.omit(unique(mRNA2name))
hela_rnaRPKM$geneName <- mRNA2name$geneName[match(hela_rnaRPKM$mRNA,mRNA2name$mRNA)]

Annot.cod$rpkm <- hela_rnaRPKM$RNAseq_HeLa_RPKM[match(Annot.cod$geneName,hela_rnaRPKM$geneName)]
Annot.cod$gro_score <- hela_rnaRPKM$GROseq_score[match(Annot.cod$geneName,hela_rnaRPKM$geneName)]
Annot.cod <- na.omit(Annot.cod)
#selection only the RPKM >0 genes list (16336)
Annot.cod.rpkm <- subset(Annot.cod,rpkm > 0)
```

in Shell:

```
computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName codingGenesAll.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 500 --maxThreshold 1000 --scoreFileName DRIP_run1_HeLa.bw DRIP_run1_TOP1.bw --outFileName "drip.allcodinggene.2cells.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "drip.allcodinggene.2cells.matrix" --outFileName "20191024.drip.allcodinggene.2cells.png" --averageType mean --colors red blue --samplesLabel HeLa shTOP1 --startLabel TSS --endLabel TTS --yMin 0.25 --yMax 2 --plotTypese --legendLocation lower-right --perGroup &
```

Fig.2d - all coding genes for DRIP & pRPA + RFD with error band Ctrl

in Shell:

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName codingGenesAll.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 500 --maxThreshold 1000 --minThreshold -100 --scoreFileName DRIP_run1_HeLa.bw pRPA_run2_HeLa.bw --outFileName "hela.drip.rpa.allcodinggene.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "hela.drip.rpa.allcodinggene.matrix" --outFileName "20191024.hela.drip.rpa.allcodinggene.png" --averageType mean --colors red black --samplesLabel HeLa-DRIP HeLa-pRPA --startLabel TSS --endLabel TTS --yMin 0 --yMax 1.75 --plotType se --legendLocation lower-right --perGroup &

RFD

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName codingGenesAllPlus.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 1000 --scoreFileName Hela.EdU.Combined_OkaSeq.RFD.bw --outFileName "RFD.allcodinggene.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "RFD.allcodinggene.matrix" --outFileName "20200225.RFD.geneOri.allcodinggene.png" --averageType mean --startLabel "TSS" --endLabel "TTS" --colors blue --plotType se --legendLocation lower-left

Fig 4g - all coding genes for DRIP & pRPA + RFD with error band shTop1

in Shell:

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName codingGenesAll.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 500 --maxThreshold 1000 --minThreshold -100 --scoreFileName DRIP_run1_TOP1.bw pRPA_run2_TOP1.bw --outFileName "top.drip.rpa.allcodinggene.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "top.drip.rpa.allcodinggene.matrix" --outFileName "top.drip.rpa.allcodinggene.png" --averageType mean --colors red black --samplesLabel

TOP1-DRIP TOP1-pRPA --startLabel TSS --endLabel TTS --yMin 0 --yMax 1.75 --plotType se --legendLocation lower-right --perGroup &

#################### Fig 4h - pairwise plot for gH2AX ####################

in R:

```
df.fiveless$level2 <- "others"
df.fiveless$level2[df.fiveless$gH2AX_run2_ASF_RPKM > df.fiveless$gH2AX_run2_HeLa_RPKM] <- "shSRSF1>Ctrl"

ggplot(df.fiveless, aes(x = log10(gH2AX_run2_HeLa_RPKM), y = log10(gH2AX_run2_ASF_RPKM))) +
geom_point(size = 2, alpha = 0.7, aes(color = level2)) +
geom_smooth(method=lm) +
scale_color_manual(values=c("black","red"))+
xlim(-1,1)+
ylim(-1,1)+
ylab("gH2AX_run2_ASF_RPKM") +
xlab("gH2AX_run2_HeLa_RPKM") +
theme_classic()+
geom_abline(intercept = 0, slope = 1, color = "black")

fit1 <- lm(gH_hela ~ gH_top, na.action=na.exclude, data = df.tt)
fit1 <- lm(gH2AX_run2_HeLa_RPKM ~ gH2AX_run2_TOP1_RPKM, na.action=na.exclude, data = df.fiveless)
R2 <- signif(summary(fit1)$adj.r.squared,5)
Pval <-signif(summary(fit1)$coef[2,4], 5)

df.fiveless$level <- "others"
df.fiveless$level[df.fiveless$gH2AX_run2_TOP1_RPKM > df.fiveless$gH2AX_run2_HeLa_RPKM] <- "shTOP1>Ctrl"

ggplot(df.fiveless, aes(x = log10(gH2AX_run2_HeLa_RPKM), y = log10(gH2AX_run2_TOP1_RPKM))) +
geom_smooth(method=lm) +
geom_point(size = 2, alpha =0.7, aes(color = level)) +
scale_color_manual(values=c("black","red"))+
xlim(-1,1)+
```

```
ylim(-1,1)+
ylab("gH2AX_run2_TOP1_RPKM") +
xlab("gH2AX_run2_HeLa_RPKM") +
theme_classic()+
geom_abline(intercept = 0, slope = 1, color = "black")
```

########## Fig.5a top25% and other genes bless signals in shTop1 +/- 2kb TTS ########

in R:

```
bless_hela <- read.table("B_HeLa_S1_L001_R1_HWGLVBCXY.bedgraph", col.names = c("chr","start","end","bless_hela"))
bless_TOP1 <- read.table("B_HeLa_shTop1_S2_L002_R1_HWGLVBCXY.bedgraph", col.names = c("chr","start","end","score"))
```

generation the +/- 2kb TTS region

in sh:

```
### top25% BLESS signals metagene profiles :
refGene_all_plus <- refGene_all[refGene_all$strand == "+", 1:5]
refGene_all_minus <- refGene_all[refGene_all$strand == "-", 1:5]

refGene_TTS2kb_plus <- data.frame(chr= refGene_all_plus$chromosome, start= refGene_all_plus$end - 2000, end= refGene_all_plus$end + 2000,geneID= refGene_all_plus$mRNA,strand = refGene_all_plus$strand)
refGene_TTS2kb_minus <- data.frame(chr= refGene_all_minus$chromosome, start= refGene_all_minus$start - 2000, end= refGene_all_minus$start + 2000,geneID= refGene_all_minus$mRNA,strand = refGene_all_minus$strand)
refGene_TTS2kb <- rbind(refGene_TTS2kb_plus,refGene_TTS2kb_minus)
refGene_TTS2kb <- refGene_TTS2kb[with(refGene_TTS2kb, order(chr,start)), ]
library(ggplot2)
library(ggpubr)
library(data.table)

setDT(refGene_TTS2kb)
setDT(bless_TOP1)
setDT(bless_hela)
setkey(refGene_TTS2kb,chr,start,end)
```

```
refG_tts_bless_TOP1 <- foverlaps(bless_TOP1,refGene_TTS2kb,type = "any",mult = "all",nomatch = 0L)
aggr_refG_bless_TOP1 <- aggregate(score ~ geneID,refG_tss_bless_TOP1, median)
quant_top1_bless <- quantile(aggr_refG_bless_TOP1$score,probs = seq(0, 1, 0.25))
aggr_refG_bless_TOP1$level <- "other"
aggr_refG_bless_TOP1$mRNA_rpkm <- df$RNAseq_HeLa_RPKM[match(aggr_refG_bless_TOP1$geneID,df$ID)]
aggr_refG_bless_TOP1[aggr_refG_bless_TOP1$score > quant_top1_bless[4],]$level <- "top25%"
aggr_refG_bless_TOP1_top25 <- aggr_refG_bless_TOP1[aggr_refG_bless_TOP1$level == "top25%",]
aggr_refG_bless_TOP1_other <- aggr_refG_bless_TOP1[aggr_refG_bless_TOP1$level == "other",]
gene_refG_bless_TOP1_top25 <- refGene_all[match(aggr_refG_bless_TOP1_top25$geneID, refGene_all$mRNA),c(1:4,10,5)]
gene_refG_bless_TOP1_other <- refGene_all[match(aggr_refG_bless_TOP1_other$geneID, refGene_all$mRNA),c(1:4,10,5)]
gene_refG_bless_TOP1 <- refGene_all[match(aggr_refG_bless_TOP1$geneID, refGene_all$mRNA),c(1:4,10,5)]

########################## heatmap bless. ##################################

computeMatrix reference-point --numberOfProcessors 8 --regionsFileName 20200325_gene_bless_top1_2kbtop25.bed 20200325_gene_bless_top1_2kbother75.bed --scoreFileName B_HeLa_S1_L001_R1_HWGLVBCXY.bw B_HeLa_shTop1_S2_L002_R1_HWGLVBCXY.bw --outFileName "heatMap.blesstop25.matrix" --beforeRegionStartLength 5000 --afterRegionStartLength 5000 --binSize 100 --missingDataAsZero --skipZeros --referencePoint TES
plotHeatmap --matrixFile "heatMap.blesstop25.matrix" --outFileName "20200403.heatMap.bless2kbtop25.png" --colorList lavender,royalblue,gold,darkorange,firebrick --whatToShow "plot, heatmap and colorbar" --refPointLabel TTS --samplesLabel "BLESS-HeLa" "BLESS-shTOP1" --regionsLabel "top 25% bless" "others" --zMax 4 --yMax 4

####### Fig.5b top25% and other genes bless signals in shTop1 +/- 2kb TTS ###########
computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName 20200325_gene_bless_top1_2kbtop25.bed 20200325_gene_bless_top1_2kbother75.bed --
```

beforeRegionStartLength 5000 --afterRegionStartLength 5000 --regionBodyLength 10000 --binSize 100 --maxThreshold 200 --scoreFileName DRIP_run1_HeLa.bw DRIP_run1_TOP1.bw --outFileName "drip.blesstop25.matrix" --missingDataAsZero --skipZeros
plotProfile --matrixFile "drip.blesstop25.matrix" --outFileName "20200403.drip.bless2kbtop25.png" --averageType mean --colors red black --samplesLabel "drip-hela" "drip-shTOP1" --regionsLabel "top25%" "others" --plotTitle " " --endLabel TTS --plotType se

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName 20200325_gene_bless_top1_2kbtop25.bed 20200325_gene_bless_top1_2kbother75.bed --beforeRegionStartLength 5000 --afterRegionStartLength 5000 --regionBodyLength 10000 --binSize 100 --minThreshold -0.7 --scoreFileName pRPA_run2_HeLa.bw pRPA_run2_TOP1.bw --outFileName "rpa.blesstop25.matrix" --missingDataAsZero --skipZeros
plotProfile --matrixFile "rpa.blesstop25.matrix" --outFileName "20200403.rpa.bless2kbtop25.png" --averageType mean --colors red black --samplesLabel "rpa-hela" "rpa-shTOP1" --regionsLabel "top25%" "others" --plotTitle " " --endLabel TTS --plotType se

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName 20200325_gene_bless_top1_2kbtop25.bed 20200325_gene_bless_top1_2kbother75.bed --beforeRegionStartLength 5000 --afterRegionStartLength 5000 --regionBodyLength 10000 --binSize 100 --minThreshold 0 --maxThreshold 10 --scoreFileName gH2AX_run2_HeLa.bw gH2AX_run2_TOP1.bw --outFileName "gH2AX.blesstop25.matrix" --missingDataAsZero --skipZeros
plotProfile --matrixFile "gH2AX.blesstop25.matrix" --outFileName "20200403.gH2AX.bless2kbtop25.png" --averageType mean --colors red black --samplesLabel "gH2AX-hela" "gH2AX-shTOP1" --regionsLabel "top25%" "others" --plotTitle " " --endLabel TTS --plotType se

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName 20200325_gene_bless_top1_2kbtop25.bed 20200325_gene_bless_top1_2kbother75.bed --beforeRegionStartLength 5000 --afterRegionStartLength 5000 --regionBodyLength 10000 --binSize 100 --scoreFileName B_HeLa_S1_L001_R1_HWGLVBCXY.bw B_HeLa_shTop1_S2_L002_R1_HWGLVBCXY.bw --outFileName "bless.blesstop25.matrix" --missingDataAsZero --skipZeros

```
plotProfile --matrixFile "bless.blesstop25.matrix" --outFileName "20200403.bless.bless2kbtop25.png" --averageType mean --colors red black --samplesLabel "bless-hela" "bless-shTOP1" --regionsLabel "top25%" "others" --plotTitle " " --endLabel TTS --plotType se


############## S.Fig.1f - mRNA RPKM in R-loop shTop1-specific genes ##############
top_geneOnly <- read.table("intersect_TOP1_DRIP_genes.bed",col.names = c("chromosome","start","end","ID","score","strand"))
hela_geneOnly <- read.table("intersect_HeLa_DRIP_genes.bed",col.names = c("chromosome","start","end","ID","score","strand"))
top_geneOnly_rpkm <- unique(df[match(df$ID,top_geneOnly$ID),])
top_geneOnly_rpkm <- top_geneOnly_rpkm[!is.na(top_geneOnly_rpkm$chromosome),]
hela_geneOnly_rpkm <- unique(df[match(df$ID,hela_geneOnly$ID),])
hela_geneOnly_rpkm <- hela_geneOnly_rpkm[!is.na(hela_geneOnly_rpkm$chromosome),]

setDT(top_geneOnly_rpkm)
setDT(hela_geneOnly_rpkm)
setkey(top_geneOnly_rpkm,chromosome,start,end,ID)
setkey(hela_geneOnly_rpkm,chromosome,start,end,ID)
top_specif <- top_geneOnly_rpkm[!hela_geneOnly_rpkm]
df.rnaHela <- data.frame(rpkm=top_specif$RNAseq_HeLa_RPKM, class="RNA_HeLa")
df.rnaASF <- data.frame(rpkm=top_specif$RNAseq_ASF_RPKM, class="RNA_ASF")
df.rnaTOP1 <- data.frame(rpkm=top_specif$RNAseq_TOP1_RPKM, class="RNA_TOP1")
df.rna.3cells.top <- rbind(df.rnaHela,df.rnaASF,df.rnaTOP1)
df.rna.3cells.top$rpkm <- log10(df.rna.3cells.top$rpkm)
df.rna.3cells.top$cell <- "shTop1 specific genes"

ggplot(data=df.rna.3cells, aes(x=class,y=rpkm)) +
geom_boxplot(aes(fill=class),size = 1,width=0.4) +
facet_wrap(~ cell, scales="free") +
scale_fill_manual(values=c("gray100", "gray75", "gray50"))+
ylim(-5,3)+
ylab("mRNA level (log10 RPKM)")+
ggtitle("transcription expression for shTop1 and shASF specific genes in three cell lines")+
stat_compare_means(comparisons = my_comparisons,label.y = c(2.8, 3))+
stat_compare_means(label = "p.signif", method = "t.test",
```

```
ref.group = "gH2AX_HeLa", hide.ns = TRUE) +
theme_light()+
theme(axis.title= element_text(size=15, color="black", face= "bold", vjust=0.5, hjust=0.5),
axis.text= element_text(size=15, color="black", face= "bold", vjust=0.5, hjust=0.5))+
theme(legend.position="top")

############################  S.Fig 1g  ################################
rf_Data <- read.csv('01_count_RPKM_and_peaks_intersect_on_refSeq_hg19.csv', sep = '\t',
header = TRUE)
Annot.cod.rpkm <- subset(Annot.cod,rpkm > 0)

for(i in c(1:22)) {
chr.name <- paste0("chr",i)
print(chr.name)
chr.length <- chr.sizes[chr.sizes$chr == chr.name,2]
print(chr.length)

list.name<- paste("listmap",sep = ".",chr.name)

Annot.cod.rpkm.chr <- Annot.cod.rpkm[Annot.cod.rpkm$chr == chr.name,]
Annot.cod.rpkm.plus <- Annot.cod.rpkm.chr[Annot.cod.rpkm.chr$strand =="+",]
Annot.cod.rpkm.minus <- Annot.cod.rpkm.chr[Annot.cod.rpkm.chr$strand == "-",]
listmap <- rep(0,chr.length)
if(nrow(intersect_codGene_chiaPET_fin_genes.chr.plus)>1){
for (i in 1:nrow(Annot.cod.rpkm.plus)){
for (j in (Annot.cod.rpkm.plus$start[i]):(Annot.cod.rpkm.plus$end[i])){
listmap[j]<- 1
}
}
}
if(nrow(intersect_codGene_chiaPET_fin_genes.chr.minus)>1){
for (i in 1:nrow(Annot.cod.rpkm.minus)){
for (j in (Annot.cod.rpkm.minus$start[i]):(Annot.cod.rpkm.minus$end[i])){
if (listmap[j]==0)listmap[j]<- 2
else listmap[j]<- 3

}
```

```
}
}
assign(list.name,listmap)

}

Annot.cod.rpkm$scoreStrand_up <- Annot.cod.rpkm$distance_up <-
Annot.cod.rpkm$scoreStrand_dw <- Annot.cod.rpkm$distance_dw <- 0
Annot.cod.rpkm <- Annot.cod.rpkm[with(Annot.cod.rpkm, order(chr,start)), ]

#################### upstream genes distance and orientation ############

for(i in c(1:22)) {
chr.name <- paste0("chr",i)
print(chr.name)
Annot.cod.rpkm.chr <- Annot.cod.rpkm[Annot.cod.rpkm$chr == chr.name,]
mx <- max(Annot.cod.rpkm.chr$end)
listmap.chr.name <- paste0("listmap",seq =".",chr.name)
listmap.chr <- get(listmap.chr.name)
for (j in 1:(nrow(Annot.cod.rpkm.chr)-1) ) {
if (Annot.cod.rpkm.chr$strand[j+1] == "+")
{
pos <- Annot.cod.rpkm.chr$start[j+1]-1
dis <-1
score <- listmap.chr[pos]
while(score==0 & pos > 1)
{
pos<-pos-1
dis <- dis+1
score <- listmap.chr[pos]
}
Annot.cod.rpkm[Annot.cod.rpkm$chr == chr.name,]$scoreStrand_up[j+1] <- score
Annot.cod.rpkm[Annot.cod.rpkm$chr == chr.name,]$distance_up[j+1] <- dis
}
if (Annot.cod.rpkm.chr$strand[j] == "-")
{
pos <- Annot.cod.rpkm.chr$end[j] +1
```

```
dis <- 1
score <- listmap.chr[pos]
while (score==0 & pos < mx) {
pos <-pos+1
dis <-dis+1
score <- listmap.chr[pos]
}
Annot.cod.rpkm[Annot.cod.rpkm$chr == chr.name,]$scoreStrand_up[j]<- score
Annot.cod.rpkm[Annot.cod.rpkm$chr == chr.name,]$distance_up[j]<-dis
}
}
}


################### downstream genes distance and orientation    ###################

for(i in c(1:22)) {
chr.name <- paste0("chr",i)
print(chr.name)
Annot.cod.rpkm.chr <- Annot.cod.rpkm[Annot.cod.rpkm$chr == chr.name,]
mx <- max(Annot.cod.rpkm.chr$end)
listmap.chr.name <- paste0("listmap",seq =".",chr.name)
listmap.chr <- get(listmap.chr.name)
for (j in 1:(nrow(Annot.cod.rpkm.chr)-1) ) {
if (Annot.cod.rpkm.chr$strand[j+1] == "-")
{
pos <- Annot.cod.rpkm.chr$start[j+1]-1
dis <-1
score <- listmap.chr[pos]
while(score==0 & pos > 1)
{
pos<-pos-1
dis <- dis+1
score <- listmap.chr[pos]
}
Annot.cod.rpkm[Annot.cod.rpkm$chr == chr.name,]$scoreStrand_dw[j+1] <- score
Annot.cod.rpkm[Annot.cod.rpkm$chr == chr.name,]$distance_dw[j+1] <- dis
```

```
}
if (Annot.cod.rpkm.chr$strand[j] == "+")
{
pos <- Annot.cod.rpkm.chr$end[j] +1
dis <- 1
score <- listmap.chr[pos]
while (score==0 & pos < mx) {
pos <-pos+1
dis <-dis +1
score <- listmap.chr[pos]
}
Annot.cod.rpkm[Annot.cod.rpkm$chr == chr.name,]$scoreStrand_dw[j]<- score
Annot.cod.rpkm[Annot.cod.rpkm$chr == chr.name,]$distance_dw[j]<-dis
}
}
}
```

```
computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName dw_AO_5kb.bed dw_AO_more5kb.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 100 --minThreshold -100 --scoreFileName pRPA_run2_HeLa.bw --outFileName "pRPA.dwAO.moreLess5kb.hela.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "pRPA.dwAO.moreLess5kb.hela.matrix" --outFileName "20200225_pRPA.dwAO.moreLess5kb.hela.png" --averageType mean --colors red blue --regionsLabel "<5kb" ">5kb" --samplesLabel " " --plotTitle "pRPA.moreLess5kb.hela" --yMin 0 --yMax 4.2 --endLabel TTS --plotType se

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName dw_AO_5kb.bed dw_AO_more5kb.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 100 --minThreshold -100 --scoreFileName pRPA_run2_TOP1.bw --outFileName "pRPA.dwAO.moreLess5kb.shTop1.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "pRPA.dwAO.moreLess5kb.shTop1.matrix" --outFileName "20200225_pRPA.dwAO.moreLess5kb.shTop1.png" --averageType mean --colors red blue --samplesLabel " " --regionsLabel "<5kb" ">5kb" --plotTitle "pRPA.moreLess5kb.shTop1" --yMin 0 --yMax 4.2 --endLabel TTS --plotType se
```

```
########################## S.Fig 1h ##########################################
computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName dw_AO_5kbRPKMmore1BothGenes.bed dw_AO_5kbRPKMOther.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 100 --minThreshold -100 --scoreFileName pRPA_run2_HeLa.bw --outFileName "pRPA.dwAO.less5kb.RNArpkm1.hela.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "pRPA.dwAO.less5kb.RNArpkm1.hela.matrix" --outFileName "20200225_pRPA.dwAO.less5kb.RNArpkm1.hela.png" --averageType mean --colors red blue --regionsLabel "<5kb-RPKM>1" "<5kb-RPKM<1" --samplesLabel " " --plotTitle "pRPA.rpkm1.less5kb.hela" --yMin 0 --yMax 4.2 --endLabel TTS --plotType se

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName dw_AO_5kbRPKMmore1BothGenes.bed dw_AO_5kbRPKMOther.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 100 --minThreshold -100 --scoreFileName pRPA_run2_TOP1.bw --outFileName "pRPA.dwAO.less5kb.RNArpkm1.shTop1.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "pRPA.dwAO.less5kb.RNArpkm1.shTop1.matrix" --outFileName "20200225_pRPA.dwAO.less5kb.RNArpkm1.shTop1.png" --averageType mean --colors red blue --samplesLabel " " --regionsLabel "<5kb-RPKM>1" "<5kb-RPKM<1" --plotTitle "pRPA.rpkm1.less5kb.shTop1" --yMin 0 --yMax 4.2 --endLabel TTS --plotType se
########################## S.Fig 1j ##########################################
aggr_drip_merge_rloop_peak_Mergenes_hela <- aggregate(i.score ~ geneName,merge_rloop_peak_Mergenes_drip_hela, median)
merge_rloop_peak_Mergenes_drip_top1 <- foverlaps(drip_TOP1,rloop_peak_genes_all,type = "any",mult = "all",nomatch = 0L)
aggr_drip_merge_rloop_peak_Mergenes_top1 <- aggregate(i.score ~ geneName,merge_rloop_peak_Mergenes_drip_top1, median)

merge_rloop_peaks_drip_hela <- foverlaps(drip_hela,merge_rloop_peaks,type = "any",mult = "all",nomatch = 0L)
aggr_drip_merge_rloop_peaks_hela <- aggregate(score ~ start,merge_rloop_peaks_drip_hela, median)
```

```
aggr_drip_rloopPeak_Mergenes <- cbind(aggr_drip_merge_rloop_peak_Mergenes_hela,aggr_drip_merge_rloop_peak_Mergenes_top1)
aggr_drip_rloopPeak_Mergenes$level <- "both"

aggr_drip_rloopPeak_Mergenes[match(rloop_peak_genes_onlyInHeLa_2$geneName,aggr_drip_rloopPeak_Mergenes$geneName),]$level <- "OnlyInHeLa"

aggr_drip_rloopPeak_Mergenes[match(rloop_peak_genes_onlyInTop1_2$geneName,aggr_drip_rloopPeak_Mergenes$geneName),]$level <- "OnlyInshTop1"



ggscatter(aggr_drip_rloopPeak_Mergenes, x = "i.score", y = "i.score_shTop1",
color = "level",
palette = c("gray", "blue","red"), shape = 20, size = 1.5,
ylab = "DRIP_Genes_with_rloopPeak_shTop1",
xlab = "DRIP_Genes_with_rloopPeak_HeLa",
title = "scatter-plot for the R-loop signals in genes with Rloop peak",
ylim= c(0, 40),
xlim= c(0, 40)

)+
geom_abline(intercept = 0, slope = 1, color = "black")



###################### S.Fig 2g ##############################################
computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName dw_AO_5kb.bed dw_AO_more5kb.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 100 --minThreshold -100 --scoreFileName pRPA_run2_HeLa.bw --outFileName "pRPA.dwAO.moreLess5kb.hela.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "pRPA.dwAO.moreLess5kb.hela.matrix" --outFileName "20200225_pRPA.dwAO.moreLess5kb.hela.png" --averageType mean --colors red blue --regionsLabel "<5kb" ">5kb" --samplesLabel " " --plotTitle "pRPA.moreLess5kb.hela" --yMin 0 --yMax 4.2 --endLabel TTS --plotType se
```

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName dw_AO_5kb.bed dw_AO_more5kb.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 100 --minThreshold -100 --scoreFileName pRPA_run2_TOP1.bw --outFileName "pRPA.dwAO.moreLess5kb.shTop1.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "pRPA.dwAO.moreLess5kb.shTop1.matrix" --outFileName "20200225_pRPA.dwAO.moreLess5kb.shTop1.png" --averageType mean --colors red blue --samplesLabel " " --regionsLabel "<5kb" ">5kb" --plotTitle "pRPA.moreLess5kb.shTop1" --yMin 0 --yMax 4.2 --endLabel TTS --plotType se

############################## S.Fig 2h ###################################
computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName dw_AO_5kbRPKMmore1BothGenes.bed dw_AO_5kbRPKMOther.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 100 --minThreshold -100 --scoreFileName pRPA_run2_HeLa.bw --outFileName "pRPA.dwAO.less5kb.RNArpkm1.hela.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "pRPA.dwAO.less5kb.RNArpkm1.hela.matrix" --outFileName "20200225_pRPA.dwAO.less5kb.RNArpkm1.hela.png" --averageType mean --colors red blue --regionsLabel "<5kb-RPKM>1" "<5kb-RPKM<1" --samplesLabel " " --plotTitle "pRPA.rpkm1.less5kb.hela" --yMin 0 --yMax 4.2 --endLabel TTS --plotType se

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName dw_AO_5kbRPKMmore1BothGenes.bed dw_AO_5kbRPKMOther.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 100 --minThreshold -100 --scoreFileName pRPA_run2_TOP1.bw --outFileName "pRPA.dwAO.less5kb.RNArpkm1.shTop1.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "pRPA.dwAO.less5kb.RNArpkm1.shTop1.matrix" --outFileName "20200225_pRPA.dwAO.less5kb.RNArpkm1.shTop1.png" --averageType mean --colors red blue --samplesLabel " " --regionsLabel "<5kb-RPKM>1" "<5kb-RPKM<1" --plotTitle "pRPA.rpkm1.less5kb.shTop1" --yMin 0 --yMax 4.2 --endLabel TTS --plotType se

################################## S.Fig 2i ################################
computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName ori_dw_1_5kb.bed ori_dw_5_20kb.bed ori_dw_20_50kb.bed --beforeRegionStartLength 10000 --

afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 100 --minThreshold -100 --scoreFileName pRPA_run2_HeLa.bw --outFileName "pRPA.dwAO.ori.distance.hela.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "pRPA.dwAO.ori.distance.hela.matrix" --outFileName "20200225_pRPA.dwAO.ori.distance.hela.png" --averageType mean --colors red black blue --regionsLabel "<5kb" "5-20kb" "20-50kb" --samplesLabel " " --plotTitle "pRPA.gene-origin.dis.hela" --yMin 0 --yMax 2.2 --endLabel TTS --plotType se

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName ori_dw_1_5kb.bed ori_dw_5_20kb.bed ori_dw_20_50kb.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 100 --minThreshold -100 --scoreFileName pRPA_run2_TOP1.bw --outFileName "pRPA.dwAO.ori.distance.shTop1.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "pRPA.dwAO.ori.distance.shTop1.matrix" --outFileName "20200225_pRPA.dwAO.ori.distance.shTop1.png" --averageType mean --colors red black blue --samplesLabel " " --regionsLabel "<5kb" "5-20kb" "20-50kb" --plotTitle "pRPA.gene-origin.dis.shTop1" --yMin 0 --yMax 2.2 --endLabel TTS --plotType se

################################## S.Fig 2j ##################################
computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName ori_dw_1_5kb_plus.bed ori_dw_5_20kb_plus.bed ori_dw_20_50kb_plus.bed ori_dw_50_up_plus.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 10000 --binSize 1000 --scoreFileName Hela.EdU.Combined_OkaSeq.RFD.bw --outFileName "ori.dwAO.ori.plus.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "ori.dwAO.ori.plus.matrix" --outFileName "ori.dwAO.ori.plus.png" --averageType mean --regionsLabel "<5kb" "5-20kb" "20-50kb" ">50kb" --startLabel "TSS" --endLabel "TTS" --legendLocation lower-left

############################## S.Fig 5b ##################################
#HeLa

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName genes_list_5kb_TES_k2.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 500 --maxThreshold 1000 --scoreFileName DRIP_run1_HeLa.bw --outFileName "drip.isma5kbTES.hela.matrix" --missingDataAsZero --skipZeros
plotProfile --matrixFile "drip.isma5kbTES.hela.matrix" --outFileName "20191024.drip.isma5kbTES.hela.png" --averageType mean --endLabel "TTS" --yMin 0 --yMax 2 --color red black --plotType se --legendLocation upper-center

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName genes_list_5kb_TES_k2.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 500 --minThreshold -100 --scoreFileName pRPA_run2_HeLa.bw --outFileName "rpa.isma5kbTES.hela.matrix" --missingDataAsZero --skipZeros
plotProfile --matrixFile "rpa.isma5kbTES.hela.matrix" --outFileName "20191024.rpa.isma5kbTES.hela.png" --averageType mean --endLabel "TTS" --yMin 0 --yMax 1.2 --color red black --plotType se --legendLocation upper-center

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName genes_list_5kb_TES_k2.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 500 --minThreshold 0 --scoreFileName gH2AX_run2_HeLa.bw --outFileName "gH2AX.isma5kbTES.hela.matrix" --missingDataAsZero --skipZeros
plotProfile --matrixFile "gH2AX.isma5kbTES.hela.matrix" --yMin 0.7 --yMax 6 --color red black --outFileName "20191024.gH2AX.isma5kbTES.hela.png" --averageType mean --endLabel "TTS" --plotType se --legendLocation upper-center

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName genes_list_5kb_TES_k2.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 500 --scoreFileName B_HeLa_S1_L001_R1_HWGLVBCXY.bw --outFileName "bless_c2_5kb_TES_k2_hela.matrix" --missingDataAsZero --skipZeros
plotProfile --matrixFile "bless_c2_5kb_TES_k2_hela.matrix" --outFileName "20191024_bless_c2_5kb_TES_k2_hela.png" --averageType mean --regionsLabel "cluster1" "cluster2" --endLabel "TTS" --yMin 0.5 --yMax 2.8 --color red black --plotType se --legendLocation upper-center

#shTop1

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName genes_list_5kb_TES_k2.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 500 --maxThreshold 1000 --scoreFileName DRIP_run1_TOP1.bw --outFileName "drip.isma5kbTES.top1.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "drip.isma5kbTES.top1.matrix" --outFileName "20191024.drip.isma5kbTES.top1.png" --averageType mean --plotType se --color red black --endLabel "TTS" --yMin 0 --yMax 2 --legendLocation upper-center

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName genes_list_5kb_TES_k2.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 500 --minThreshold -100 --scoreFileName pRPA_run2_TOP1.bw --outFileName "rpa.isma5kbTES.top1.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "rpa.isma5kbTES.top1.matrix" --outFileName "20191024.rpa.isma5kbTES.top1.png" --averageType mean --plotType se --color red black --endLabel "TTS" --yMin 0 --yMax 1.2 --legendLocation upper-center

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName genes_list_5kb_TES_k2.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 500 --minThreshold 1 --scoreFileName gH2AX_run2_TOP1.bw --outFileName "gH2AX.isma5kbTES.top1.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "gH2AX.isma5kbTES.top1.matrix" --yMin 0.7 --yMax 6 --color red black --plotType se --outFileName "20191024.gH2AX.isma5kbTES.top1.png" --averageType mean --endLabel "TTS" --legendLocation lower-center

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName genes_list_5kb_TES_k2.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 500 --scoreFileName B_HeLa_shTop1_S2_L002_R1_HWGLVBCXY.bw --outFileName "bless_c2_5kb_TES_k2_top1.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "bless_c2_5kb_TES_k2_top1.matrix" --outFileName "20191024_bless_c2_5kb_TES_k2_top1.png" --averageType mean --regionsLabel "cluster1"

"cluster2" --endLabel "TTS" --yMin 0.5 --yMax 2.8 --color red black --plotType se --legendLocation upper-center

############################## S.Fig 5c ##################################

in R:

```
library(ggplot2)
library(ggpubr)
library(data.table)

setDT(refGene_TSS2kb)
setDT(bless_TOP1)
setDT(bless_hela)
setkey(refGene_TSS2kb,chr,start,end)
refG_tss_bless_TOP1 <- foverlaps(bless_TOP1,refGene_TSS2kb,type = "any",mult = "all",nomatch = 0L)
aggr_refG_bless_TOP1 <- aggregate(score ~ geneID,refG_tss_bless_TOP1, median)
refG_tss_bless_hela <- foverlaps(bless_hela,refGene_TSS2kb,type = "any",mult = "all",nomatch = 0L)
aggr_refG_bless_hela <- aggregate(score ~ geneID,refG_tss_bless_hela, median)
# +/- 2kb TSS resgion clusters
quant_top1_bless <- quantile(aggr_refG_bless_TOP1$score,probs = seq(0, 1, 0.25))
#  0%       25%       50%       75%
# 0.000000  0.000000  1.213596  3.222010
# 100%
# 433.245000
#
aggr_refG_bless_TOP1$level <- "other"
aggr_refG_bless_TOP1[aggr_refG_bless_TOP1$score > quant_top1_bless[4],]$level <- "top25%"
aggr_refG_bless_TOP1_top25 <- aggr_refG_bless_TOP1[aggr_refG_bless_TOP1$level == "top25%",]
aggr_refG_bless_TOP1_other <- aggr_refG_bless_TOP1[aggr_refG_bless_TOP1$level == "other",]
gene_refG_bless_TOP1_top25 <- refGene_all[match(aggr_refG_bless_TOP1_top25$geneID, refGene_all$mRNA),c(1:4,10,5)]
# 8785
```

```
gene_refG_bless_TOP1_other <- refGene_all[match(aggr_refG_bless_TOP1_other$geneID,
refGene_all$mRNA),c(1:4,10,5)]
# 26467

# top25% of delta(shTop1- hela)  +/- 2kb
aggr_refG_bless_TOP1$delta <- aggr_refG_bless_TOP1$score -
aggr_refG_bless_hela$score

quant_top1_bless_delta<- quantile(aggr_refG_bless_TOP1$delta,probs = seq(0, 1, 0.25))
#       0%        25%        50%
# -179.180250   -1.682957    0.000000
# 75%        100%
#   1.888770   31.495600

aggr_refG_bless_TOP1$level_delta <- "other"

aggr_refG_bless_TOP1[aggr_refG_bless_TOP1$delta >
quant_top1_bless_delta[4],]$level_delta <- "deltop25"
aggr_refG_bless_TOP1_deltop25 <-
aggr_refG_bless_TOP1[aggr_refG_bless_TOP1$level_delta == "deltop25",]
aggr_refG_bless_TOP1_del75 <-
aggr_refG_bless_TOP1[aggr_refG_bless_TOP1$level_delta == "other",]

gene_refG_bless_TOP1_deltop25 <-
refGene_all[match(aggr_refG_bless_TOP1_deltop25$geneID,
refGene_all$mRNA),c(1:4,10,5)]
#8812
gene_refG_bless_TOP1_del75 <- refGene_all[match(aggr_refG_bless_TOP1_del75$geneID,
refGene_all$mRNA),c(1:4,10,5)]
#26440

write.table(gene_refG_bless_TOP1_top25,  "20200325_gene_bless_top1_tss2kbtop25.bed",
quote = FALSE, sep = "\t", col.names=F, row.names=F)
write.table(gene_refG_bless_TOP1_other,  "20200325_gene_bless_top1_tss2kb75.bed",
quote = FALSE, sep = "\t", col.names=F, row.names=F)
```

```
write.table(gene_refG_bless_TOP1_deltop25,
"20200325_gene_bless_top1_tss2kbdel25.bed", quote = FALSE, sep = "\t", col.names=F,
row.names=F)
write.table(gene_refG_bless_TOP1_del75,  "20200325_gene_bless_top1_tss2kbdel75.bed",
quote = FALSE, sep = "\t", col.names=F, row.names=F)

##############################  mRNA P1593.  ##############################@
mRNA_hela_tts2kb <- aggr_refG_bless_TOP1[, c(1,3,5,6)]
colnames(mRNA_hela_tts2kb)[4] <- "mRNA_rpkm"
mRNA_hela_tts2kb$cell <- "ctrl"
colnames(mRNA_top1_tts2kb)[4] <- "mRNA_rpkm"
mRNA_top1_tts2kb <- aggr_refG_bless_TOP1[, c(1,3,5,7)]
mRNA_top1_tts2kb$cell <- "shTop1"
mRNA_all_tts2kb <- rbind(mRNA_hela_tts2kb,mRNA_top1_tts2kb)

my_comparisons <- c("ctrl","shTop1")
ggplot(data=mRNA_all_tts2kb, aes(x=cell,y=log2(mRNA_rpkm))) +
#geom_violin(width=0.6, position=position_dodge(0.75), bw=1.5)+
geom_boxplot(aes(fill=cell),size = 1,width=0.4) +
facet_wrap(~ level, scales="free") +
scale_fill_manual(values=c("gray100", "gray75", "gray50"))+
#ylim(-5,6)+
ylab("mRNA_rpkm")+
ggtitle("mRNA signal distribution for the genes in +/-2kb TTS top25%")+
stat_compare_means(comparisons = my_comparisons,label.y = c(10, 10.2))+
stat_compare_means(label = "p.signif", method = "t.test",
hide.ns = TRUE,label.y = 10.4) +
theme_light()+
theme(axis.title= element_text(size=15, color="black", face= "bold", vjust=0.5, hjust=0.5),
axis.text= element_text(size=15, color="black", face= "bold", vjust=0.5, hjust=0.5))+
theme(legend.position="top")

ggplot(data=mRNA_all_tts2kb, aes(x=cell,y=log2(mRNA_rpkm))) +
#geom_violin(width=0.6, position=position_dodge(0.75), bw=1.5)+
geom_boxplot(aes(fill=cell),size = 1,width=0.4) +
facet_wrap(~ level_delta, scales="free") +
scale_fill_manual(values=c("gray100", "gray75", "gray50"))+
```

```
#ylim(-5,6)+
ylab("mRNA_rpkm")+
ggtitle("mRNA signal distribution for the genes in +/-2kb TTS delta top25%")+
stat_compare_means(comparisons = my_comparisons,label.y = c(10, 10.2))+
stat_compare_means(label = "p.signif", method = "t.test",
hide.ns = TRUE,label.y = 10.4) +
theme_light()+
theme(axis.title= element_text(size=15, color="black", face= "bold", vjust=0.5, hjust=0.5),
axis.text= element_text(size=15, color="black", face= "bold", vjust=0.5, hjust=0.5))+
theme(legend.position="top")
```

```
computeMatrix reference-point --numberOfProcessors 8 --regionsFileName 20200325_gene_bless_top1_tss2kbtop25.bed 20200325_gene_bless_top1_tss2kb75.bed --scoreFileName B_HeLa_S1_L001_R1_HWGLVBCXY.bw B_HeLa_shTop1_S2_L002_R1_HWGLVBCXY.bw --outFileName "heatMap.blesstop25tss.matrix" --beforeRegionStartLength 5000 --afterRegionStartLength 5000 --binSize 100 --missingDataAsZero --skipZeros --referencePoint TSS

plotHeatmap --matrixFile "heatMap.blesstop25tss.matrix" --outFileName "20200407.heatMap.bless2kbtop25tss.png" --colorList lavender,royalblue,gold,darkorange,firebrick --whatToShow "plot, heatmap and colorbar" --refPointLabel TSS --samplesLabel "BLESS-HeLa" "BLESS-shTOP1" --regionsLabel "top 25% bless" "others" --zMax 4 --yMax 4


########################## S.Fig 5d ###################################
computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName 20200325_gene_bless_top1_tss2kbtop25.bed 20200325_gene_bless_top1_tss2kb75.bed --beforeRegionStartLength 5000 --afterRegionStartLength 5000 --regionBodyLength 10000 --binSize 100 --maxThreshold 200 --scoreFileName DRIP_run1_HeLa.bw DRIP_run1_TOP1.bw --outFileName "drip.blesstop25tss.matrix" --missingDataAsZero --skipZeros
plotProfile --matrixFile "drip.blesstop25tss.matrix" --outFileName "20200407.drip.bless2kbtop25tss.png" --averageType mean --colors red black --samplesLabel "drip-hela" "drip-shTOP1" --regionsLabel "top25%tss" "others" --plotTitle " " --endLabel TTS --plotType se
```

```
computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName 20200325_gene_bless_top1_tss2kbtop25.bed 20200325_gene_bless_top1_tss2kb75.bed --beforeRegionStartLength 5000 --afterRegionStartLength 5000 --regionBodyLength 10000 --binSize 100 --minThreshold -5 --scoreFileName pRPA_run2_HeLa.bw pRPA_run2_TOP1.bw --outFileName "rpa.blesstop25tss.matrix" --missingDataAsZero --skipZeros
plotProfile --matrixFile "rpa.blesstop25tss.matrix" --outFileName "20200407.rpa.bless2kbtop25tss.png" --averageType mean --colors red black --samplesLabel "rpa-hela" "rpa-shTOP1" --regionsLabel "top25%" "others" --plotTitle " " --endLabel TTS --plotType se

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName 20200325_gene_bless_top1_tss2kbtop25.bed 20200325_gene_bless_top1_tss2kb75.bed --beforeRegionStartLength 5000 --afterRegionStartLength 5000 --regionBodyLength 10000 --binSize 100 --scoreFileName B_HeLa_S1_L001_R1_HWGLVBCXY.bw B_HeLa_shTop1_S2_L002_R1_HWGLVBCXY.bw --outFileName "bless.blesstop25tss.matrix" --missingDataAsZero --skipZeros
plotProfile --matrixFile "bless.blesstop25tss.matrix" --outFileName "20200407.bless.bless2kbtop25tss.png" --averageType mean --colors red black --samplesLabel "bless-hela" "bless-shTOP1" --regionsLabel "top25%" "others" --plotTitle " " --endLabel TTS --plotType se


############################  S.Fig 6a  ######################################
my_comparisons <- list( c("cluster1", "cluster2") )
ggplot(data=aggr_drip_2cells, aes(x=condition,y=log2(score))) +
geom_boxplot(aes(fill=condition),size = 1,width=0.4) +
facet_wrap(~ cell, scales="free") +
scale_fill_manual(values=c("gray100", "gray75", "gray50"))+
ylim(-5,6)+
ylab("log2(DRIP)")+
ggtitle("DRIP signal distribution for the genes in 2 clusters in 2 cells")+
stat_compare_means(comparisons = my_comparisons,label.y = c(5.5, 5.7))+
stat_compare_means(label = "p.signif", method = "t.test",
hide.ns = TRUE,label.y = 5.9) +
theme_light()+
theme(axis.title= element_text(size=15, color="black", face= "bold", vjust=0.5, hjust=0.5),
axis.text= element_text(size=15, color="black", face= "bold", vjust=0.5, hjust=0.5))+
```

```
theme(legend.position="top")


ggplot(data=aggr_rpa_2cells, aes(x=condition,y=log2(score))) +
geom_boxplot(aes(fill=condition),size = 1,width=0.4) +
facet_wrap(~ cell, scales="free") +
scale_fill_manual(values=c("gray100", "gray75", "gray50"))+
ylim(-5,6)+
ylab("log2(pRPA)")+
ggtitle("pRPA signal distribution for the genes in 2 clusters in 2 cells")+
stat_compare_means(comparisons = my_comparisons,label.y = c(5.5, 5.7))+
stat_compare_means(label = "p.signif", method = "t.test",
hide.ns = TRUE,label.y = 5.9) +
theme_light()+
theme(axis.title= element_text(size=15, color="black", face= "bold", vjust=0.5, hjust=0.5),
axis.text= element_text(size=15, color="black", face= "bold", vjust=0.5, hjust=0.5))+
theme(legend.position="top")

ggplot(data=aggr_gH2AX_2cells, aes(x=condition,y=log2(score))) +
geom_boxplot(aes(fill=condition),size = 1,width=0.4) +
facet_wrap(~ cell, scales="free") +
scale_fill_manual(values=c("gray100", "gray75", "gray50"))+
ylim(-5,6)+
ylab("log2(gH2AX)")+
ggtitle("gH2AX signal distribution for the genes in 2 clusters in 2 cells")+
stat_compare_means(comparisons = my_comparisons,label.y = c(5.5, 5.7))+
stat_compare_means(label = "p.signif", method = "t.test",
hide.ns = TRUE,label.y = 5.9) +
theme_light()+
theme(axis.title= element_text(size=15, color="black", face= "bold", vjust=0.5, hjust=0.5),
axis.text= element_text(size=15, color="black", face= "bold", vjust=0.5, hjust=0.5))+
theme(legend.position="top")

ggplot(data=aggr_bless_2cells, aes(x=condition,y=log2(score))) +
geom_boxplot(aes(fill=condition),size = 1,width=0.4) +
facet_wrap(~ cell, scales="free") +
scale_fill_manual(values=c("gray100", "gray75", "gray50"))+
```

```
ylim(-5,6)+
ylab("log2(bless)")+
ggtitle("BLESS signal distribution for the genes in 2 clusters in 2 cells")+
stat_compare_means(comparisons = my_comparisons,label.y = c(5.5, 5.7))+
stat_compare_means(label = "p.signif", method = "t.test",hide.ns = TRUE,label.y = 5.9) +
theme_light()+
theme(axis.title= element_text(size=15, color="black", face= "bold", vjust=0.5, hjust=0.5),
axis.text= element_text(size=15, color="black", face= "bold", vjust=0.5, hjust=0.5))+
theme(legend.position="top")
```

```
################################## S.Fig 6b ###################################
computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName 20190725_c1_dwHO_disLess5kb.bed 20190725_c1_dwHO_disMore5kb.bed 20190725_c2_dwHO_disLess5kb.bed 20190725_c2_dwHO_disMore5kb.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 100 --maxThreshold 1000 --scoreFileName DRIP_run1_HeLa.bw --outFileName "DRIP.2c.HOless5kb.hela.matrix" --missingDataAsZero --skipZeros
plotProfile --matrixFile "DRIP.2c.HOless5kb.hela.matrix" --outFileName "20190725.DRIP.2c.HOless5kb.hela.png" --averageType mean --colors red green blue black --regionsLabel "C1-HOless5kb" "C1-HOmore5kb" "C2-HOless5kb" "C2-HOmore5kb" --yMin 0 --yMax 6 --endLabel TTS --legendLocation upper-left
computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName 20190725_c1_dwHO_disLess5kb.bed 20190725_c1_dwHO_disMore5kb.bed 20190725_c2_dwHO_disLess5kb.bed 20190725_c2_dwHO_disMore5kb.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 100 --maxThreshold 1000 --scoreFileName DRIP_run1_TOP1.bw --outFileName "DRIP.2c.HOless5kb.top1.matrix" --missingDataAsZero --skipZeros
plotProfile --matrixFile "DRIP.2c.HOless5kb.top1.matrix" --outFileName "20190725.DRIP.2c.HOless5kb.top1.png" --averageType mean --colors red green blue black --regionsLabel "C1-HOless5kb" "C1-HOmore5kb" "C2-HOless5kb" "C2-HOmore5kb" --yMin 0 --yMax 6 --endLabel TTS --legendLocation upper-left

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName 20190725_c1_dwHO_disLess5kb.bed 20190725_c1_dwHO_disMore5kb.bed 20190725_c2_dwHO_disLess5kb.bed 20190725_c2_dwHO_disMore5kb.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 -
```

-binSize 100 --minThreshold -100 --scoreFileName pRPA_run2_HeLa.bw --outFileName "pRPA.2c.HOless5kb.hela.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "pRPA.2c.HOless5kb.hela.matrix" --outFileName "20190725.pRPA.2c.HOless5kb.hela.png" --averageType mean --colors red green blue black --regionsLabel "C1-HOless5kb" "C1-HOmore5kb" "C2-HOless5kb" "C2-HOmore5kb" --yMin 0 --yMax 2.5 --endLabel TTS --legendLocation upper-left

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName 20190725_c1_dwHO_disLess5kb.bed 20190725_c1_dwHO_disMore5kb.bed 20190725_c2_dwHO_disLess5kb.bed 20190725_c2_dwHO_disMore5kb.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 100 --minThreshold -100 --scoreFileName pRPA_run2_TOP1.bw --outFileName "pRPA.2c.HOless5kb.top1.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "pRPA.2c.HOless5kb.top1.matrix" --outFileName "20190725.pRPA.2c.HOless5kb.top1.png" --averageType mean --colors red green blue black --regionsLabel "C1-HOless5kb" "C1-HOmore5kb" "C2-HOless5kb" "C2-HOmore5kb" --yMin 0 --yMax 2.5 --endLabel TTS --legendLocation upper-left

############################ S.Fig 6c ####################################

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName 20190726_c1_dwCD_disLess5kb.bed 20190726_c1_dwCD_disMore5kb.bed 20190726_c2_dwCD_disLess5kb.bed 20190726_c2_dwCD_disMore5kb.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 100 --maxThreshold 1000 --scoreFileName DRIP_run1_HeLa.bw --outFileName "DRIP.2c.CDless5kb.hela.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "DRIP.2c.CDless5kb.hela.matrix" --outFileName "20190725.DRIP.2c.CDless5kb.hela.png" --averageType mean --colors red green blue black --regionsLabel "C1-CDless5kb" "C1-CDmore5kb" "C2-CDless5kb" "C2-CDmore5kb" --yMin 0 --yMax 6 --endLabel TTS --legendLocation upper-left

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName 20190726_c1_dwCD_disLess5kb.bed 20190726_c1_dwCD_disMore5kb.bed 20190726_c2_dwCD_disLess5kb.bed 20190726_c2_dwCD_disMore5kb.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 100 --maxThreshold 1000 --scoreFileName DRIP_run1_TOP1.bw --outFileName "DRIP.2c.CDless5kb.top1.matrix" --missingDataAsZero --skipZeros

plotProfile --matrixFile "DRIP.2c.CDless5kb.top1.matrix" --outFileName "20190725.DRIP.2c.CDless5kb.top1.png" --averageType mean --colors red green blue black

--regionsLabel "C1-CDless5kb" "C1-CDmore5kb" "C2-CDless5kb" "C2-CDmore5kb" --yMin 0 --yMax 6 --endLabel TTS --legendLocation upper-left

computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName 20190726_c1_dwCD_disLess5kb.bed 20190726_c1_dwCD_disMore5kb.bed 20190726_c2_dwCD_disLess5kb.bed 20190726_c2_dwCD_disMore5kb.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 100 --minThreshold -100 --scoreFileName pRPA_run2_HeLa.bw --outFileName "pRPA.2c.CDless5kb.hela.matrix" --missingDataAsZero --skipZeros
plotProfile --matrixFile "pRPA.2c.CDless5kb.hela.matrix" --outFileName "20190725.pRPA.2c.CDless5kb.hela.png" --averageType mean --colors red green blue black --regionsLabel "C1-CDless5kb" "C1-CDmore5kb" "C2-CDless5kb" "C2-CDmore5kb" --yMin 0 --yMax 2.5 --endLabel TTS --legendLocation upper-left
computeMatrix scale-regions --numberOfProcessors 8 --regionsFileName 20190726_c1_dwCD_disLess5kb.bed 20190726_c1_dwCD_disMore5kb.bed 20190726_c2_dwCD_disLess5kb.bed 20190726_c2_dwCD_disMore5kb.bed --beforeRegionStartLength 10000 --afterRegionStartLength 10000 --regionBodyLength 20000 --binSize 100 --minThreshold -100 --scoreFileName pRPA_run2_TOP1.bw --outFileName "pRPA.2c.CDless5kb.top1.matrix" --missingDataAsZero --skipZeros
plotProfile --matrixFile "pRPA.2c.CDless5kb.top1.matrix" --outFileName "20190725.pRPA.2c.CDless5kb.top1.png" --averageType mean --colors red green blue black --regionsLabel "C1-CDless5kb" "C1-CDmore5kb" "C2-CDless5kb" "C2-CDmore5kb" --yMin 0 --yMax 2.5 --endLabel TTS --legendLocation upper-left

RÉSUMÉ

Le programme de réplication de l'ADN est fréquemment mis à l'épreuve par un stress endogène et exogène, et le stress de réplication induit par l'oncogène est un moteur majeur de la progression tumorale. L'un de ces stress réplicatifs est le conflit transcription-réplication (TRC). Un autre obstacle potentiel à la progression de la réplication est R-loop. Cependant, le mécanisme sur la façon dont les R-loops sont impliquées dans la régulation du TRC et la stabilité génomique est encore mal connu. Pendant ma thèse, j'ai développé une méthode bio-informatique pour étudier la direction des fourches de réplication et détecter les zones d'initiation. Nous l'avons appliqué avec succès pour étudier le TRC lié à R-loop et avons observé que les R-loops enrichies au niveau des sites de terminaison de la transcription (TTS) de gènes hautement exprimés montrent un niveau plus élevé de TRC frontal tandis que la fourche de réplication s'arrêtant à ces TTS empêche le TRC frontal et maintient l'intégrité du génome d'une manière dépendante de TOP1.

MOTS CLÉS

Réplication de l'ADN, directionalité de la fourche de réplication, conflit entre réplication et transcription, R-loops, instabilité génomique, analyse des données multi-omiques

ABSTRACT

The DNA replication program is frequently challenged by endogenous and exogenous stress, and the oncogene-induced replication stress is a major driver of tumor progression. One of these replicative stresses is transcription-replication conflict (TRC). Another potential obstacle for the replication progression is R-loop. However, the mechanism about how R-loops are involved in the regulation of TRC and genomic stability is still poor known. During my Ph.D. study, I have developed a bioinformatics method to study the replication fork direction and to detect the initiation zones. We successfully applied it to investigate the R-loop related TRC and observed that R-loops enriched at the transcription termination sites (TTSs) of highly expressed genes show a higher level of head-on TRC meanwhile replication fork pausing at these TTSs prevents head-on TRC and maintains genome integrity in a TOP1-dependent manner.

KEYWORDS

DNA replication, replication fork direction, conflicts between replication and transcription, R-loop, genomic instability, multi-omics data analysis