

"CHROMATIN LANDSCAPE, TRANSCRIPTION AND DNA REPLICATION ORIGIN POSITIONING"

NEREA MURUGARREN GARRIDO

ID [redacted]

A thesis submitted to the University of Birmingham for the degree of

METABOLISM AND SYSTEMS RESEARCH – MSc by Research

Lead supervisor Dr. Ildem Akerman / Co-supervisor Dr. Marco Saponaro

Institute of Metabolism & Systems Research

College of Medical and Dental Sciences

University of Birmingham

May 2022

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

ABSTRACT

DNA replication initiation is made possible by the assembly of pre- replication complexes (Pre-RCs) at genomic locations called DNA replication initiation sites (origins) during G1-phase. Although there are no conserved characteristics of metazoan origins on which Pre-RCs form, previous findings suggest that origins are strongly associated with open, active chromatin regions of the genome such as promoters, enhancers, and active histone marks. However, over a third of transcriptionally silent genes associate with DNA replication initiation activity in mESC cells, and most of these genes are bound and repressed by the Polycomb repressive complex 2 (PRC2) through the repressive H3K27me3 mark. This is the strongest observed association of an epigenetic regulator element with replication origin activity. Thus, I wished to ask whether Polycomb-mediated gene repression plays a direct role in the recruitment of DNA replication origins activity to silent genes. I comprehensively characterized the consequences of PRC2 catalytic activity (EZH2 subunit) depletion on the activity of DNA replication origins in mESC through a genome-wide evaluation. My findings suggest that absence of EZH2 results in increased DNA replication initiation activity at EZH2-bound sites. Interestingly, the increase in DNA replication initiation activity does not correlate with transcriptional de-repression or acquisition of the activating marks H3K4me3 and H3K27ac but loss of the repressive H3K27me3 mark leading to increased accessibility of chromatin.

DNA replication initiation and transcriptional initiation frequently co-localize at gene promoters. Numerous studies have examined the impact of transcription (and chromatin) and its ability to recruit DNA replication activity. In the final part (Part II) of this thesis, I began to investigate what impact the localization of DNA replication initiation may have on transcription. I used a mEF cell line in which a DNA replication origin is rendered inactive. My experiments in synchronized cells suggest that Pre-RC formation impedes transcription in a transient, time-specific manner during G1-phase of the cell cycle.

Key words Bioinformatics, cell cycle, chromatin, DNA replication initiation, DNA replication origins, genomics, Polycomb proteins, transcription

ACKNOWLEDGEMENTS

I thank my parents (Plácido and Nuria) for their emotional and financial support. Thank you for encouraging me to move out to England on my own and for always supporting me.

I acknowledge the University of Birmingham, in special the College of Medical and Dental Sciences staff, for academic support and paperwork guidance. I recognise the University of Birmingham Flow Cytometry and Genomics facilities and research technicians. I also acknowledge the University of Birmingham's BlueBEAR HPC service, which has constituted the main computing platform in this thesis.

I acknowledge the Institute of Metabolism & Systems Research (IMSR) facilities and workspace (314 lab). I acknowledge Dr. Paul Alexander Foster (Post-graduate Research Lead from the IMSR) and Viktorija Ziabliceva (Research and Knowledge Transfer from MDS) for academic guidance. I thank the IMSR staff and technicians, principal investigators and academic colleagues that supported my research and helped me through my master's.

I acknowledge Prof. Marcel Méchali and his group, the Replication and Genomic Dynamics Lab, for their constructive collaborative study with us. I appreciate the discussions and the trust in my work in bioinformatics.

I thank the Pancreatic Beta Cell Genomics Lab for in-lab and out-lab support. I thank the BSc student Sam Brain, for her valuable performance in the wet-lab experiments. I thank the postdoctoral researcher Dr. Fiona Docherty for unceasing support and teaching on her research on stem cell-derived beta cells.

I thank my co-supervisor Dr. Marco Saponaro for providing protocols and advice.

Finally, I thank Asst. Prof. Ildem Akerman for the great supervision, patient guidance and encouragement she has provided throughout my time as her master's student. I have been extremely fortunate to learn from her talent, wisdom and knowledge, specifically on DNA replication and transcription, bioinformatics and molecular biology. I certainly encountered my interest in DNA replication research through Ildem and I am taking with me many skills that will greatly impact my career.

TABLE OF CONTENTS

List of Figures and Tables 6

Abbreviations 7

1. INTRODUCTION 10

1.1. DNA replication 10
1.1.1. DNA replication initiation: origins 11
1.1.1.1. Biochemistry of origins replication: licensing and firing 11
1.1.1.2. DNA replication origins and flexibility of origin usage 12
1.1.1.3. Characteristics of origin sequences 14

PART I: Impact of Polycomb-mediated repression on DNA replication initiation 17
1.2. Chromatin environment and epigenetic marks 17
1.2.1. Chromatin features that affect origin activity 18
1.2.1.1. Polycomb repressive complexes (PRC) 19

PART II: Impact of DNA replication origin positioning on transcription 21
1.3. Transcription and transcription-related elements 21
1.3.1. Gene regulatory elements 22
1.3.1.1. Promoters 22
1.3.1.2. Enhancers 23

1.4. DNA replication initiation and transcription 23
1.4.1. Replication timing and regulation of transcription 23
1.4.2. Implications of replication origin dysregulation on transcription 24
1.4.3. Genes transcriptionally investigated 25
1.4.3.1. Retinoic acid-induced 1 (Rai1) gene 25
1.4.3.2. Phosphatidylethanolamine-N-methyltransferase (Pemt) gene 26
1.4.3.3. Sterol regulatory element-binding transcription factor 1 (Srebf-1) gene 26

2. HYPOTHESIS, OBJECTIVES & BRIEF EXPERIMENTAL PLAN 27

2.1. PART I 27
2.2. PART II 27

3. MATERIALS & METHODS 28

3.1. Tissue culture 28
3.1.1. Cell sources and culture 28
3.1.2. Cryopreservation and thawing 29

3.2. Cell synchronization and sorting 29
3.2.1. Cell synchronization and release from serum starvation 29
3.2.2. Sample preparation, staining and fluorescence-activated cell sorting 30

3.3. Local gene expression evaluation 30
3.3.1. RNA extraction for quantitative PCR 30
3.3.2. cDNA synthesis for quantitative PCR (reverse transcription) 30
3.3.3. Primer design and quality controls 31

3.3.4. Quantitative PCR 32

3.4. Sequencing 33
3.4.1. RNA extraction, and sequencing 33
3.4.2. RNA-primed short nascent strand (SNS) isolation, and sequencing 33
3.4.3. Chromatin extraction and immunoprecipitation (ChIP), and sequencing 34
3.4.4. Antibodies used in this study 35

3.5. Bioinformatics analysis pipeline 35
3.5.1. Data accession 36
3.5.2. Alignment and quality control of raw sequencing data 37
3.5.3. Peak calling 37
3.5.4. Quantification and differential peak analysis 38
3.5.5. Differential gene expression analysis 38
3.5.6. Gene ontology analysis 38
3.5.7. BEDTools 38
3.5.8. Data presentation 39
3.5.9. Data sources and availability 39

3.6. Statistical analysis 39

4. RESULTS 40

4.1. PART I 40
4.1.1. Polycomb-repressed promoters harbour DNA replication origins 40
4.1.2. Loss of EZH2 catalytic activity results in increased DNA replication initiation 42
4.1.3. Replication initiation and transcriptional de-repression in $Ezh2^{\Delta/\Delta}$ cells 45
4.1.4. Increased origin activity in $Ezh2^{\Delta/\Delta}$ cells correlates with EZH2 binding sites 47
4.1.5. Upregulation in origin activity is linked to changes in chromatin landscape 48
4.1.6. Upregulation in origin activity is linked to changes in chromatin topological states 52

4.2. PART II 56
4.2.1. The deletion of an OGRE/G4 element led to a less active replication origin 56
4.2.2. Reduced Rai1 gene expression in asynchronous $3T3^{\Delta G4}$ cells 57
4.2.3. The G-quadruplex motif act as a positive cis-regulatory element 58
4.2.4. Pre-RC formation interferes with adjacent gene expression 59

5. DISCUSSION & CONCLUSIONS 62

5.1. PART I: Re-shaped chromatin environment recruits DNA replication initiation upon loss-of-function of EZH2 62
5.1.1. Limitations and future work 64

5.2. PART II: Pre-RC formation hampers transcription from Rai1 promoter 65
5.2.1. Limitations and future work 68

6. PUBLICATION OUTCOMES 70

Bibliography 71

SUPPLEMENTARY 80

List of Figures and Tables

Figures

Figure 1. Replication origin initiation biochemistry and SNS-Seq 14
Figure 2. Metazoan replication origins characteristics 17
Figure 3. Polycomb-mediated transcription repression 20
Figure 4. Approach followed in this thesis to test the impact of local gene expression upon deletion of a G4 element 26
Figure 5. Overview of the bioinformatics pipeline followed in this thesis 35
Figure 6. Polycomb-repressed promoters harbour DNA replication origins 41
Figure 7. Loss of EZH2 catalytic activity results in increased DNA replication initiation .. 44
Figure 8. Replication initiation and transcriptional de-repression in $Ezh2^{\Delta/\Delta}$ cells 46
Figure 9. Increased origin activity in $Ezh2^{\Delta/\Delta}$ cells correlates with EZH2 binding sites 48
Figure 10. EZH2 inactivation triggers a loss of H3K27me3 mark and an upregulation of DNA replication initiation activity 50
Figure 11. EZH2 inactivation triggers a spreading of new upregulated H3K27ac peaks . 52
Figure 12. EZH2 inactivation triggers changes in the chromatin topological states 55
Figure 13. Gene expression levels of Rai1, Pemt and Srebf-1 in AS, G0-48h and G0 synchronized $3T3^{WT}$ and $3T3^{\Delta G4}$ cells 58
Figure 14. FACS analysis and normalized gene expression of $3T3^{\Delta G4}$ cells post-release 60
Figure 15. Normalized gene expression of $3T3^{\Delta G4}$ cells at different stages of the Pre-RC formation 61
Figure 16. Model suggested in this thesis by which PRC2 recruits DNA replication initiation activity 64
Figure 17. Model suggested in this thesis by which Pre-RC formation inhibits nearby transcription 68
Supplementary Figure 1. $Ezh2^{\Delta/\Delta}$ cell line genomic mutation and wet-lab control experiments 81
Supplementary Figure 2. $3T3^{\Delta G4}$ cell line genomic mutation 82
Supplementary Figure 3. Control read pile up plots for M. Méchali Lab and EZH2 binding sites datasets 83
Supplementary Figure 4. Control read pile up plots for Z. Gu et al. datasets 83

Tables

Table 1. Primer sequences designed with supporting information 32
Table 2. Sources for the publicly available data utilized in this thesis 36
Supplementary Table 1. Number of peaks for each dataset 84

Abbreviations

DNA	Deoxyribonucleic acid
S-phase	Synthesis phase of the cell cycle
Cryo-EM	Cryogenic electron microscopy
NGS	Next-generation sequencing
S. cerevisiae	*Saccharomyces cerevisiae*
D. melanogaster	*Drosophila Melanogaster*
mESC	Mouse embryonic stem cells
mEF	Mouse embryonic fibroblasts
IS	Initiation site
Kb	Kilobase
Origin	DNA replication initiation site
G1 and G2-phase	Growing phase 1 and 2 of the cell cycle
M-phase	Mitotic phase of the cell cycle
Pre-RC	Pre- replication complex
ORC 1 and 2	Origin recognition complex 1 and 2
ATP	Adenosine triphosphate
CDC6	Cell division cycle 6
CDT1	Chromatin Licensing and DNA Replication Factor 1
MCM2-7	Mini-chromosome maintenance 2-7 subunits
dsDNA	Double-stranded DNA
Pre-IC	Pre- initiation complex
IC	Initiation complex
Mcm10	Minichromosome Maintenance 10 Replication Initiation Factor
Cdc45	Cell division cycle 45
GINS	Sld5, Psf1, Psf2, and Psf3 complex
CMGM	Replicative helicase, Cdc45, Mcm2-7, GINS, Mcm10
RPA	Replication protein A
ssDNA	Single-stranded DNA
Pol	Polymerase
dNTPs	Deoxyribose nucleotide triphosphate
RNA	Ribonucleic acid
bp	Base pair
ATR	Ataxia-Telangiectasia and Rad3-related kinase
ATM	Ataxia-Telangiectasia mutated kinase
CHK1 and CHK2	Checkpoint kinases 1 and 2
OK-Seq	Okazaki sequencing
INI-Seq	Initiation site sequencing
BrdU	5-bromo-2'-deoxyuridine
dUTP	Deoxyuridine triphosphate
EdU-Seq	5-ethynyl-2’-deoxyuridicine sequencing
ChIP-Seq	Chromatin Immunoprecipitation sequencing
SNS-Seq	RNA-primed small nascent strands isolation and sequencing
FDR	False discovery rate
NS	Nascent strands of DNA
λ-Exo	λ-Exonuclease

ACS	Autonomous consensus sequences
ARS	Autonomously replicated sequences
TSS	Transcription start site
TAD	Topologically associating domain
H3K27ac	Histone 3 acetylation in lysine 27
H3K27me3	Histone 3 tri-methylation in lysine 27
CpG	5'—C—phosphate—G—3'
CGI	CpG islands
GC-rich	Guanine Cytosine -rich
AT-rich	Adenine Thymine -rich
OGRE	Origin G-rich Repeated Element
G4	G-quadruplex
OGRE/G4	Origin G-rich Repeated Element / G-quadruplex
RIF1	Replication timing regulatory factor 1
HAT	Histone acetyltransferase
HMTase	Histone methyltransferase
H3K4me3	Histone 3 tri-methylation in lysine 4
MLL	Mixed-lineage leukemia complex
ATAC-Seq	Sequencing Assay for Transposase-Accessible Chromatin
PCR	Polymerase chain reaction
PcG	Polycomb group
PRC1 and 2	Polycomb repressive complex 1 and 2
EZH 1 and 2	Enhancer of Zeste 1 and 2
EED	Embryonic Ectoderm Development
SUZ12	Suppressor of Zeste 12
SET	Su(var)3-9, Enhancer-of-Zeste and Trihorax
CTCF	CCCTC-binding factor
ncRNA	Non-coding RNA
BRE	B recognition element
TBP	TATA-binding protein
HTA1-HTB1	Histone H2A.1 and H2B.1 proteins
Ori1	Origin 1
Rai1	Retinoic acid-induced 1 gene
Pemt	Phosphatidylethanolamine-N-methyltransferase gene
Srebf-1	Sterol regulatory element-binding transcription factor 1 gene
PTLS	Potocki-Lupski synchrome
SMS	Smith-Magenis syndrome
CLOCK	Circadian locomotor output cycles kaput
RNA-Seq	RNA sequencing
DMEM	Dulbecco's Modified Eagle Medium
FBS	Fetal Bovine Serum
LIF	Leukemia Inhibitory Factor
Tmx	Tamoxifen
$3T3^{\Delta G4}$	3T3 G4 knockout cells
$3T3^{WT}$	3T3 Wild-type cells
(D)-PBS	(Dulbecco's)- phosphate-buffered saline
DMSO	Dimethyl sulfoxide

G0	G0-phase of the cell cycle
AS	Asynchronous
FACS	Fluorescence activated cell sorting
RNase	Ribonuclease
cDNA	Complementary DNA
RT	Reverse transcription
Q-PCR	Quantitative- PCR
DTT	Dithiothreitol
Tm	Melting temperature
Actb	Beta-actin
DNase	Deoxyribonuclease
$Ezh2^{+/+}$	Wild-type mESC cells
$Ezh2^{\Delta/\Delta}$	EZH2 SET domain- knockout mESC cells
RIN	RNA integrity number
PCNA	Proliferating cell nuclear antigen
OCT4	Octamer-binding transcription factor 4
HRP	Horseradish Peroxidase
IgG	Immunoglobulin G
CPT	Camptothecin
HPC	High-performance computing
HTC	High-throughput computing
LSK	Hematopoietic stem cells (Lin^{-} $Sca1^{+}$ $c\text{-}Kit^{+}$)
BS	Binding sites
TPM	Transcripts per million
mHox A	Homeobox A cluster
TMM	Trimmered Mean of M-values
DGE	Differential gene expression
CNS	Central nervous system
SpikChIP	Spike-in ChIP-Seq
$ORC1^{fl/fl}$	3T3 ORC1 knockout cells

1. INTRODUCTION

1.1. DNA replication

DNA replication is the biological process by which a deoxyribonucleic acid (DNA) molecule is copied from the parental DNA molecule and synthesized. It is a fundamental biochemical reaction that occurs in all living organisms as a part of cell division, and it confers the basis for the genetic inheritance of all genomes. Coordination and regulation of the DNA replication machinery are necessary to ensure the correct distribution of the genetic information while avoiding the accumulation of DNA damage and mutations, which can lead to genome instability with severe consequences for the cell and the organism [1-3]. For this reason, the eukaryotic cell has evolved to a fine and established order of events and components to accomplish fidelity of DNA replication during the synthesis phase (S-phase) of the interphase throughout the lifetime of an organism.

Since the discovery of the basis for DNA replication in 1953-1957 by several outstanding scientists, such as Rosalind Franklin, James Watson, Francis Crick, Matthew Meselson and Franklin Stahl, the understanding of DNA replication has been addressed from different perspectives, including bulk biochemistry of purified proteins *in vitro* [4-7], structural biology analysis of purified replication proteins using Cryo-EM or X-ray diffraction [8-10], or the direct visualization and kinetics evaluation of *in vitro* interaction of DNA and proteins through biophysics [11-13]. Lately, genome-wide approaches involving next-generation sequencing (NGS) have gained interest in the study of DNA replication [14-19]. The use of omics in the study of DNA replication can provide relevant insights behind DNA replication dysregulations that can underlie many human diseases, such as diabetes.

The use of model organisms such as *Saccharomyces cerevisiae* (*S. cerevisiae*, generally referred to as budding yeast) [20], the *Drosophila Melanogaster* (*D. melanogaster*, common fruit fly) [21] and the bacteria species *Escherichia coli* (*E. coli*) [22] provided a large fraction of our knowledge in DNA replication and conferred the basis for the study of metazoan replication, which is now focused on mammalians (mouse and human) [18, 23] and plants (*Arabidopsis thaliana*) [24]. In my thesis, I will focus on the study of replication in *Mus musculus* (mouse). I will use modified cell lines for mouse embryonic stem cells (mESC) and mouse embryonic fibroblasts (mEF).

1.1.1. DNA replication initiation: origins

Replication of the large and evolutionary complex genomes of metazoans starts at thousands (~ 30,0000-100,000 regions for mammals) of DNA replication initiation sites (IS), termed replication origins, at each cell division during S-phase [24-27]. The term “replication origins” was first described in 1968 when Huberman and Riggs studied autoradiographies of DNA and found that DNA replication started at multiple sites in the genome, spaced at intervals of 20-400 Kb, which then progressed bi-directionally at a rate of 2-3 Kb/min [28]. Eukaryotic origin regions contain the DNA site that is bound by pre- replication complexes (Pre-RCs), a set of proteins that are assembled at replication origins, and a downstream region for DNA synthesis initiation to start, where the DNA unwinds and the DNA polymerase binds [25, 26]. The replication of origins occurs as a two-step mechanism (Figure 1A, reproduced from [26]) to avoid re-replication of DNA and DNA damage intermediates [5, 29]. The two subsequent steps are the assembly of the Pre-RC at DNA (a reaction known as “licensing”) and the activation of DNA synthesis (a reaction known as “firing”) [14]. DNA replication origin licensing and firing occur at different time points during the cell cycle and are mutually exclusive, ensuring that the genome is not re-replicated [14, 23, 26, 29].

1.1.1.1. Biochemistry of origins replication: licensing and firing

Licensing and activation of origins have different biochemical components [26, 29]. Figure 1A shows an overview of the two-step replication reaction. Licensing occurs during the G1-phase of the cell cycle when Pre-RC components sequentially load to potential replication origin site. Importantly, ORC (origin recognition complex, in yellow) is the central component in this process, which binds the origin DNA in the first place. ORC is a multi-subunit complex, ORC1-6, that binds to DNA and directs the loading of Pre-RCs in an ATP-dependent manner [30, 31]. Although some components of ORC, such as ORC1, are bound to potential origins during the entirely of the cell cycle, the ATPase activity only activates in late mitosis and early G1, avoiding Pre-RC loading out of G1-phase. Following the binding of ORC, CDC6 and CDT1 proteins are also bound to the origin in G1-phase [32]. The last step of the licensing reaction is the loading of two rings of the replicative helicase, the MCM2-7 (mini-chromosome maintenance) complex [11, 20, 32]. The eukaryotic replicative helicase MCM2-7 is a heterohexameric ATPase protein complex that encircles the double-stranded DNA molecule (dsDNA) at origins and once activated, is responsible for the unwinding (or melting) of the dsDNA for the replication fork to advance during DNA synthesis [6, 33, 34]. Pre-RCs are converted into pre- initiation complexes (Pre-IC) at the G1-S phase transition with the addition

of other crucial proteins. The two-tiered MCM2-7 rings bind six additional accessory factors-Mcm10, Cdc45 and the GINS heterotetrametric complex (Sld5, Psf1, Psf2, and Psf3), and confer the activated helicase CMGM complex [11, 33, 35-37]. The replication competent Pre-IC is formed with the final addition of the replicative polymerases [14, 23]. During G1-S phase transition, the assembly of replicative proteins is controlled by two cell-cycle-regulated kinases (DDK and CDK), that phosphorylate several components [14, 23].

The final stage of DNA replication, which happens in S-phase, is the “firing” step. When the cells enter S-phase, the replication components bound to the DNA replication origins make up the IC, or replisome [14, 23]. During helicase activation, a.k.a. “firing”, the two helicase hexamers divide into two CMGM complexes, which are oriented in a head-to-head disposition [34] and convert dsDNA to single-stranded DNA (dsDNA), moving past one another and forming two replication forks emerging from the activated origin [11]. Several proteins associate with replication units and assist the progression of the replication fork, such as RPA, which binds to ssDNA and stabilizes the unwound strands [38]. Importantly, DNA synthesis by the DNA polymerase Pol ε (leading strand) can only start from RNA primers, short strands of RNA which are synthesized by an enzyme called primase (RNA polymerase), which are then continued by the incorporation of dNTPs (DNA nucleotides) [39]. Once settled, DNA synthesis is initiated by the coordinated DNA strands unwinding by CMGM and DNA synthesis by Pol ε (primed by short RNA oligos made by the DNA primase), resulting in replication forks growing from 5’ to 3’ bi-directionally from the origins [40]. The duplication of the lagging strand is carried out by the DNA polymerase Pol δ by discontinuous, short, Okazaki DNA fragments (150-200 bp) also primed from RNA and ligated together by the DNA ligase [14, 23].

1.1.1.2. DNA replication origins and flexibility of origin usage

When cells enter S-phase, many potential licensed origins contain functional bound replisome components and are ready to “fire” [14, 25, 26, 29]. However, only a third of licensed origins are activated in a defined temporal and coordinated order at a given cell cycle [14, 25, 26, 41]. Moreover, as shown in Figure 1A, within a cluster of replicons (also called replicon unit, 50-120 Kb in metazoans), the adjacent origins to the activated ones are silenced by checkpoint kinases (ATR, ATM CHK1 and CHK2) to ensure correct DNA replication initiation within a region [26]. The stochasticity of origin activation and flexibility of choice for which origins are destined to be fired are considered key aspects of DNA replication robustness and adaptation to both environmental and intracellular stress [26, 42]. However, despite extensive efforts, the molecular understanding of space and time of DNA replication origin activation in eukaryotes

is still not comprehensively known. To understand the flexibility in the choice of origins within different cells and differentiation states, as well as general study of DNA replication, genomic approaches are gaining interest.

Several approaches have been described to map origins in eukaryotic cells using NGS (reviewed in [23]). Some of the reported techniques are (i) the sequencing of the Okazaki fragment (OK-Seq) [18, 43], which are only present in the lagging replicating strand. This technique that directly measures the genome-wide replication fork directionality and infers the positions of origins from mixed cell populations; (ii) initiation site sequencing (INI-Seq) [19], in which newly synthesized DNA is labeled with BrdU-dUTP at the beginning of S-phase, and then immunoprecipitated *in vitro*; (iii) EdU-Seq, a method for mapping early S-phase origins which involves *in vitro* staining of newly synthesized DNA with the thymidine nucleotide analog 5-ethynyl-2'-deoxyuridicine (EdU) and sequencing [18, 44, 45]; (iv) chromatin immunoprecipitation and sequencing (ChIP-Seq) of the origin recognition complex 1 or 2 (ORC1 and ORC2) [46, 47], the protein complex that binds to origins at the licensing stage in the Pre-RCs formation (G1-phase); and (v) RNA-primed small nascent strands isolation and sequencing (SNS-Seq), which is the origins mapping technique that is utilized in this thesis to study DNA replication initiation. Importantly, SNS-Seq is a highly used method that minimizes false-positive rates (FPR), allowing accurate characterization of human and mouse origins with high resolution. SNS-Seq was first described in 2012 by E. Besnard et al. and is based on the isolation of 0.5-2 Kb newly synthesized ssDNA molecules (present in active origins), which are termed nascent strands (NS) [17]. The isolation of 0.5-2 Kb fragments allows the exclusion of Okazaki fragments and large fragments which are not suitable for NGS. As shown in Figure 1B (reproduced from [16]), the main principle of SNS-Seq consists of enrichment of nascent strands by extraction of total DNA, denaturation and separation of NS based on size followed by digestion of the fragments with the enzyme λ-Exonuclease (λ-Exo – enriching for RNA-primed NS DNA), coupled with NGS [16, 17]. Genome-wide replication profiling with SNS-Seq data allowed the discovery of many features about metazoan origins [16, 48-50].

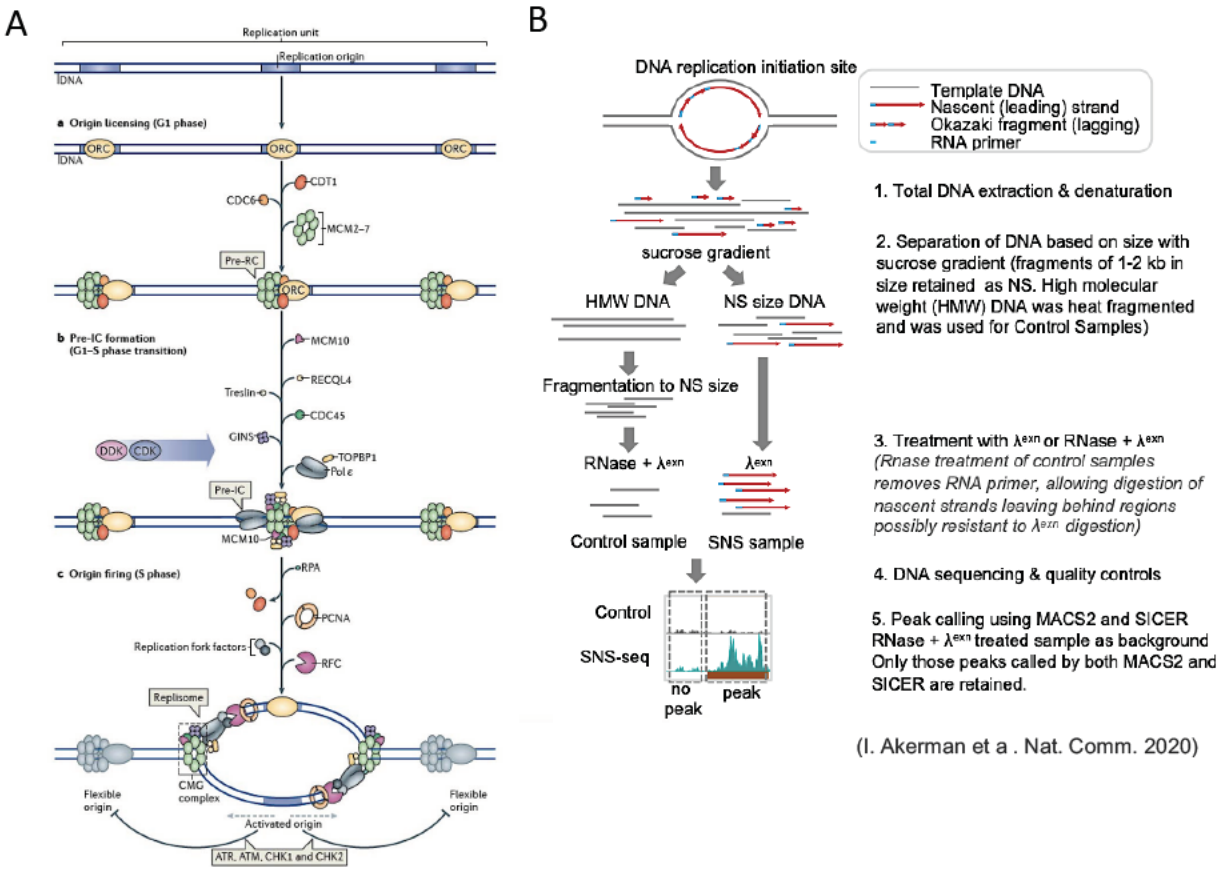


Figure 1. Replication origin initiation biochemistry and SNS-Seq. (A) Format on and act vat on of DNA rep cat on or g ns. The f gure shows a rep cat on un t w th three potent a rep cat on or g ns and how these are censed by the recogn t on of the or g n recogn t on comp ex (ORC) and the ORC-d rected format on of Pre-rep cat on comp exes (Pre-RCs) dur ng G1-phase, a react on ca ed or g n cens ng (a). Pre-RCs are transformed nto Pre- n t at on comp exes (Pre-ICs) dur ng G1-S phase trans t on by the add t on of other cruc a prote ns, nc ud ng Cdc45 and GINS. The rep cat on competent Pre-IC s formed w th the f na add t on of the rep cat ve po ymerases (b). Or g n act vat on and f r ng occurs dur ng S-phase, where the doub e-stranded DNA (dsDNA) mo ecu e unw nds fo ow ng the act vat on of the MCM comp ex to form two rep cat on forks, wh ch advance n a 5 to 3 d rect ona ty by the coord nated funct on of the he case (CMG comp ex) and the DNA po ymerase, conferr ng the funct ona rep some comp ex (DNA e ongat on) (c). As shown n the f gure, many factors regu ate th s process, nc ud ng severa ce cyc e k nases (ATR, ATM, CHK1 and CHK2). Reproduced from [26]. **(B)** Fu exper menta SNS-Seq procedure, nc ud ng contro s. The ma n pr nc p e of RNA-pr med sma nascent strands so at on and sequenc ng (SNS-Seq) cons sts of the extract on of tota DNA, the denaturat on and separat on of NS based on the s ze on a sucrose grad ent and the d gest on of the fragments w th the enzyme λ-Exonuc ease (λ-Exo), coup ed w th NGS. Reproduced from [16]. Further nformat on n the text.

1.1.1.3. Characteristics of origin sequences

Despite the availability of many genome-wide methods to map origins, strong discrepancies have been described between the different techniques, such as the quantification, position,

sequence features and broadness of the detected origins [17, 23, 51]. In *S. cerevisiae* genome-wide analysis, the yeast ORC recognizes and assembles at autonomous consensus sequences (ACS) found in autonomously replicated sequences (ARS) in the genome, which are regions of about 12 bp enriched in A/T bases [20, 52, 53]. Distinctively, EdU-Seq in mammalian cells suggests that origin regions are enriched in long AT-rich sequences (>20 bp), similar to the *S. cerevisiae* [45]. However, unlike *S. cerevisiae*, mapping of replication origins in fruit fly and mammalian cells using SNS-Seq or INI-Seq has shown that the ORC complex has no consensus sequence specificity but associates with G-rich DNA sequences [16, 21, 47]. This is consistent with the previous biochemistry assays using purified proteins and DNA *in vitro* [5, 7]. Besides discrepancies between mapping methodologies, several recent genome-wide studies revealed common clues on the contribution of genetic and epigenetic characteristics in the vicinity of eukaryotic DNA replication origins, although none can be considered universal predictive features [8, 15, 16, 49, 50]. Origins identified with ORC1 and ORC2 binding sites (ChIP-Seq), as well as INI-Seq and OK-Seq, revealed chromatin marks, sequence signatures and other features which are comparable with the characteristics found in origins identified by SNS-Seq [46, 47]. These techniques reported that the positioning of mammalian DNA replication origins in the genome is not arbitrary; instead, origins (and ORC binding sites) are strongly associated with nucleosome-free active chromatin (euchromatin) regions of the genome such as transcriptional start sites (TSS), TAD borders, enhancers and active histone chromatin marks (H3K27ac and H3K4me3, features discussed below) [14, 16, 19, 25, 43, 46, 48]. However, not all origins are associated with open chromatin regions. OK-Seq, INI-Seq and SNS-Seq allowed the definition of two types of origins from a kinetics perspective; early-firing (early S-phase) origins (representing ~32%) of all origins, were associated with active transcription start sites (TSS) and enhancer marks at large domains of open chromatin (sensible to DNase I digestion) [54]. On the other hand, late-firing (late S-phase) origins represented regions of inactive and chromatin, gene-poor regions and repressive marks [18, 19, 54, 55].

The sequence content is another associated characteristic with metazoan origins under debate. One of the sequences origins associated is CpG islands (CGIs), which consist of highly conserved regions of approximately 500bp to 2 Kb in length with a high G and C nucleotide content (GC-rich regions), typically found at transcription promoters (Figure 2A, reproduced from [16]) [8, 56]. CGIs sites, located at 250-300 bp upstream of IS to which the DNA polymerase machinery is recruited, were confirmed by SNS-Seq, INI-Seq and OK-Seq for more than 60% of active origins in fly, mouse and human cells, contrary to what articles performing EdU-Seq described (presence of AT-rich regions in origins) [8, 14, 16, 19, 21, 25, 41, 48, 49, 57, 58]. Although replication origins frequently localize to promoters and that

feature could explain the preferential association of origins with CGIs [56], non-promoter CGIs are also associated with replication origins [8]. Thus, CGIs may influence the origin positioning and/or activation [59]. Methylation of CGIs may function in regulating replication origin timing as origins co-localizing to unmethylated CGIs have been associated with early replication in comparison to methylated CGIs [23, 25, 54]. However, the majority of origins do not contain a CGI, which suggests this element in itself cannot be the only determinant of origins sites. Consistently, the presence of CG-rich elements was described as a predictive origin parameter by M. Méchali Lab [16, 60-62] and others [63]. When found nearby origin sites, CGIs are considered Origin G-rich Repeated Elements (OGREs), structures that can topologically form G-quadruplex motifs (OGRE/G4) [41, 62, 64]. G-quadruplexes are helical structures of nucleic acids rich in guanine, stabilized by Hoogsteen hydrogen bonding, forming square planar structures called guanine tetrads [65]. These structures are often found at the end of chromosomes (telomeric regions) and in transcriptional regulatory regions, both in prokaryotes [66] and eukaryotes [67]. OGREs are found in 80-90% of the origins in human and mouse cells, 250-300 bp upstream the DNA synthesis initiation site in nucleosome-free regions known to be Pre-RC sites (Figure 2B) [41, 49, 62]. Evidence by P. Prorok et al. suggests that the presence of G-quadruplexes increases the efficiency of origins [41]. It has been suggested that ORC may “scan” the genome through unspecific DNA binding and get “trapped” at such G-quadruplexes [60, 64]. Recent studies on the telomere-associated protein RIF1 (Replication timing regulatory factor 1), also highlighted the role of G4s in origin firing and origin efficiency [41, 68-70]. RIF1 tunes the timing of origins by delaying Pre-RC activation until the late S-phase in heterochromatin/ late-replicating regions through DNA loop formation by association with multiple G-quadruplex sequences [41, 68-70]. P. Prorok et al. also evidenced in mouse cells that (i) the deletion of OGRE/G4 led to a decrease in origin replication activity and (ii) introducing an OGRE/G4 element in an origin-free region led to the establishment of a new functional origin [41]. However, it should be noted that while ~0.5-1.5 M G4 sequences are present in the human genome, only a part of them host DNA replication origins [41].

Given the association of genomic active regions, promoter regions and GC-rich elements with DNA replication initiation, there is reason to study DNA replication initiation in a chromatin context, where chromatin modification and epigenetic inheritance occur.

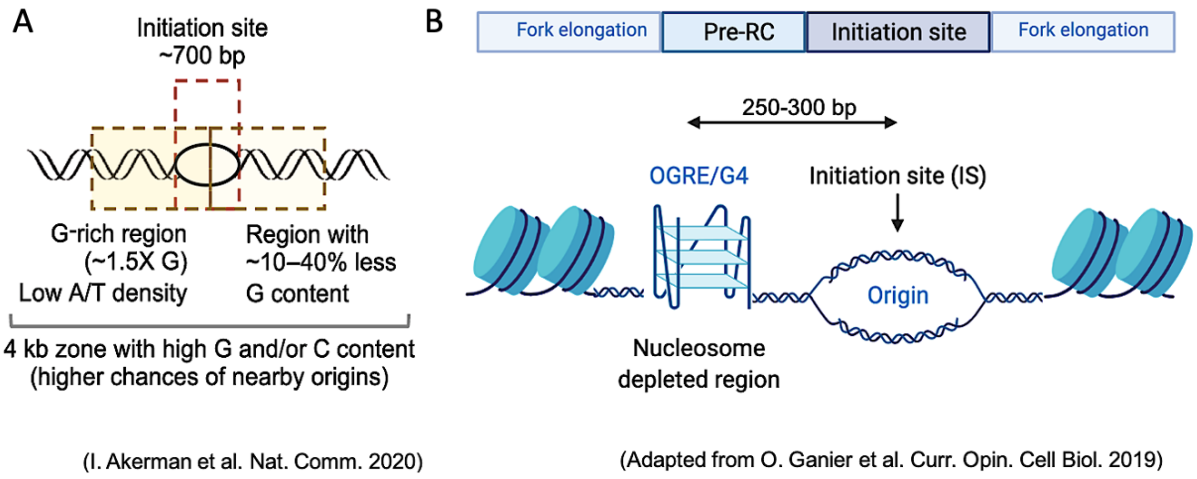


Figure 2. The characteristics of metazoan replication origins. (A) Schemat c f gure show ng the propert es of pred cted or g ns. GC-r ch reg ons have h gher changes of conta n or g ns at nearby chromat n. Reproduced from [16]. **(B)** Scheme of a metazoan rep cat on or g n. The Pre-RC oads to a sequence ocated upstream of the rep cat on n t at on s te (IS). OGRE/G4 mot fs are common y found n Pre-RC b nd ng s tes, 250-300 bp upstream of the IS. Adapted from [23]. F gure made w th B orender.com

PART I: Impact of Polycomb-mediated repression on DNA replication initiation

1.2. Chromatin environment and epigenetic marks

The concept of "epigenetics" was first introduced in 1942 by the embryologist Conrad H. Waddington, who defined it as the complex of developmental processes between the genotype and phenotype [71]. Epigenetics now involves the study of the changes in chromatin modifications and gene expression rather than alteration of the genome itself. Changes in the chromatin modifications within the nucleus are important to prevent DNA physical and chemical damage, to regulate gene expression and replication and to facilitate proper segregation of the chromosomes during cell division, besides other functions [72, 73]. These changes in chromatin modifications are translated into a differential chromosome topological state and density, in which the nucleosomes compact the DNA in different levels of chromatin organization [73-75]. Moreover, the chromatin accessibility influences the metabolism of DNA, principally DNA replication and transcription. For instance, those regions that are transcriptionally more active are typically less compacted, as open chromatin allows a better accessibility to transcription factors [73-75]. By contrast, those regions containing transcriptionally inactive genes tend to be more compacted in dense chromatin (heterochromatin) [76]. The relaxed/dense ratio in the cell is regulated by epigenetic modifications of DNA or histones, generally via acetylation and methylation [72, 73, 77]. Many of these regulatory modifications are found on histone H3 [73].

1.2.1. Chromatin features that affect origin activity

Specific histone marks have been linked to replication origin activity, for example, histone acetylation [78, 79]. Chromatin acetylation by histone acetyltransferases (HATs) leads to a higher content of euchromatin in the cell which, in turn, can increase gene expression [72]. The most studied histone acetylation is the mono-acetylation of the Lysine 27 residue of histone H3 (H3K27ac), which is found as discrete regions at both proximal and distant regions of transcription start sites (TSS), and most specifically at transcriptional enhancers. From *S. cerevisiae* to mammalian cells, active origins have been associated with open chromatin [26, 79]. In *D. melanogaster* and *S. cerevisiae* hyperacetylated histones coincide with ORC binding sites and ARS consensus sequences, respectively [78, 79]. On the other hand, histone methylation marks, deposited by histone methyltransferases (HMTases), can lead to higher contents of open or dense chromatin depending on the methylated residue. For instance, the tri-methylation of the Lysine 4 residue of histone H3 (H3K4me3) is found as discrete regions in promoter regions and is associated with a positive regulation of transcription by promoting the action of histone acetylases and nucleosome remodeling enzymes, making the chromatin more accessible [80-82]. However, the tri-methylation of the Lysine 27 residue of histone H3 (H3K27me3) is found as large regions/domains and is associated with the downregulation of the transcription of nearby genes via chromatin compaction [83, 84]. H3K27me3 has been associated with active origins, specifically those that “fire” in mid-S phase [26, 63]. H3K27me3 mark is deposited by the Polycomb repressive complex (PRC), which will be introduced in the next section of this report. Another regulatory epigenetic feature is the presence of antagonistic histone modification marks, what is called bivalent chromatin regions, which are regions containing both H3K4me3 (activating) and H3K27me3 (repressive) histone modification marks. H3K4me3 is an activating mark deposited by the MLL (mixed-lineage leukemia) complex [85, 86]. Bivalent chromatin domains are generally associated with promoters of developmentally related, repressed genes in pluripotent stem cells or multipotent progenitors [87-91]. This mark is thought to be present in more than 2,000 genes in human ESC [89] and ~4000 in mouse [92]. Bivalent domains confer a silencing environment while rendering the possibility of fast activation of the repressed genes, a very important feature in undifferentiated cell states. Upon ESC cell differentiation, most bivalent regions resolve to differential univalent states, depending on the levels of activation [91]. What is little explored is how these regions relate to DNA replication initiation.

The dynamics of the landscape of histone post-translational modification marks and chromatin states in the flexibility and activation of origins are not yet well described [93]. To address the investigation of those interrelated biological processes, the use of sequencing techniques is

widely used in the field. As revealed by chromatin immunoprecipitating and sequencing (ChIP-Seq), the changes in specific histone marks can be positioned genome-wide. Complementing the changes in histone epigenetic marks, the chromatin topological states (and chromatin accessibility) of the epigenome can also be mapped genome-wide using several sequencing techniques, such as ATAC-Seq. ATAC-Seq stands for Assay for Transposase-Accessible Chromatin using sequencing and is a technique that maps regions of loosely packaged chromatin through the insertion of sequencing adapters by a hyperactive Tn5 transposase mutant into open chromatin (a process called “tagmentation”), which is then followed by amplification by PCR and sequencing [75, 94]. In this thesis, I will show the effects of the changes in chromatin marks and states in DNA replication initiation through ChIP-Seq and ATAC-Seq, combined with SNS-Seq (origin mapping). In specific, I will discuss the role of Polycomb repressive complexes (PRC) in the recruitment of DNA replication initiation.

1.2.1.1. Polycomb repressive complexes (PRC)

One of the most important epigenetic regulatory systems is the evolutionary conserved Polycomb (PcG) repressive complexes (PRC). Correct functioning of PRC has been linked to the successful embryonic body and brain development through epigenetic regulation of differentiation genes and processes, such as X-inactivation during female development [89, 95] and the re-modeling of chromosomal chromatin during cell mitosis [96]. Polycomb proteins are recruited to open chromatin by recognition of promoters containing CpG islands (Figure 3, reproduced from [97]), which bind and regulate the epigenome by depositing chromatin repressive marks, thereby facilitating an increase in the cell dense/relaxed chromatin ratio and silencing the CGI promoters [56, 98-100]. As shown in Figure 3, there are two catalytically major complexes of PcG proteins: Polycomb Repressive Complex 1 (PRC1) and 2 (PRC2), which are inter-regulated. PRC2 is composed of 4 core subunits. The PRC2 catalytic protein is Enhancer of Zeste 2 (EZH2), or its homolog EZH1, which functions as a histone methyltransferase (HMTase) [97, 98, 101-103]. EZH2 (or EZH1) catalytically adds tri-methylations at the 27th lysine residue of the tail of histone H3 [98-100]. The other 3 subunits are Embryonic Ectoderm Development (EED) [104], Suppressor of Zeste 12 (SUZ12) and RbAp46/48. [105]. Depletion and/or mutation of PRC2 subunits leads to severe developmental defects at the early stages of mouse embryogenesis [106, 107]. PRC2 and can recruit PRC1 through the recognition of H3K27me3 by the CBX subunit of PRC1. PRC1 ubiquitinylates the 119th lysine residue of histone H2A through the E3 ubiquitin-protein ligase RING1A/B, strengthening gene transcription repression [97, 108]. Genome-wide mapping of replication origins in mouse embryonic stem cells (mESC) has shown that a fraction of activated origins

is associated with Polycomb-bound and repressed genes [16]. However, the role of PRC in DNA replication initiation is unclear.

Genome-wide studies of histone methylation in ESC cells led to the unexpected discovery that PRC sites not only overlapped with the repressive H3K27me3 mark but also the activating H3K4me3 mark [89-91]. As stated before, regions containing both H3K27me3 and H3K4me3, which are termed “bivalent domains”, are important in developmental regulation because they keep regions silent while leaving them “poised” for alternate fates [89-91]. Our understanding of PRC regulation of bivalent domains in embryonic states is still not well grounded. Even though, a remarkable genome-wide study by M. Ku et al. found that simple genomic codes involving locations, sizes, and content of CpG islands could predict the localization of PRC1 and 2, as well as bivalent domains in ESC cells [89]. The role of CpG islands in PRC recruitment is still under investigation.

Several studies performed in the Lab of M. Méchali suggest that the presence of a CpG island at the promoter of a silenced gene can also be associated with DNA replication initiation activity [8, 16, 61]. Such finding was an insight into a possible relationship between CpG islands, PRC-mediated gene repression and DNA replication initiation activity.

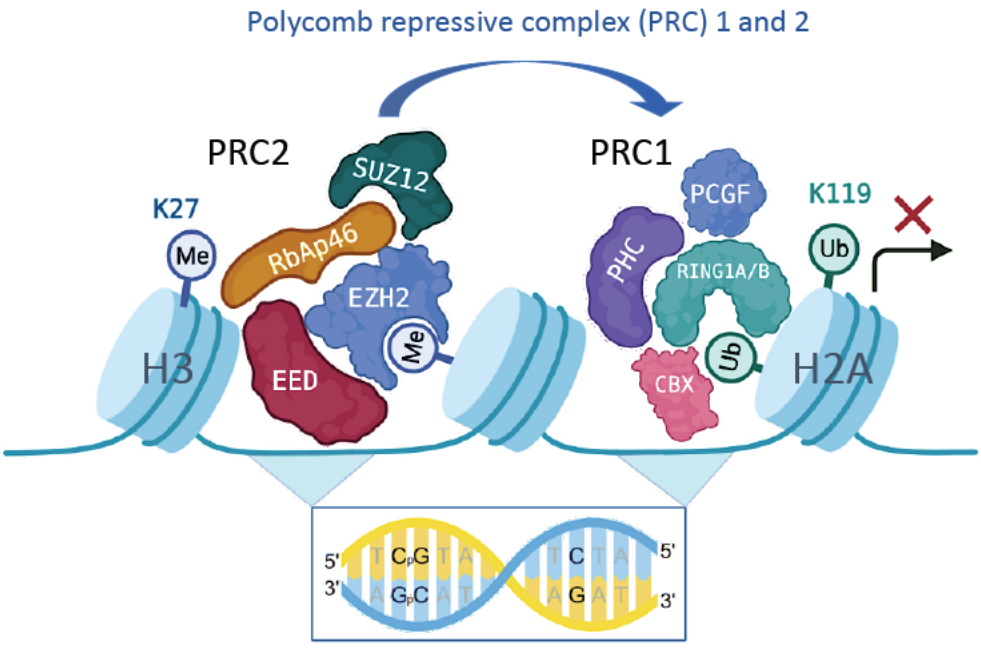


Figure 3. Polycomb-mediated transcriptional repression. Po ycomb repress ve comp exes (PRC) are recru ted to CpG s ands. The Po ycomb repress ve comp ex 2 (PRC2) conta ns four core prote ns: EZH2 (Enhancer-of-Zeste homo og 2, the cata yt c subun t), EED (Embryon c Ectoderm Deve opment), SUZ12 (z nc f nger) and RbAp46/48

(Retinoblastoma Binding Protein, the PRC2 scaffold subunit). PRC2 deposits tri-methylations at the 27th lysine residue of the tail of histone H3 (H3K27me3) through the function of the EZH2 subunit. The H3K27me3 mark represses PRC2-target genes. PRC2 recruits PRC1 through the CBX protein from Polycomb repressive complex 1 (PRC1), which recognizes H3K27me3. PRC1 contains the catalytic E3 ubiquitin-protein ligase protein RING1A/B, which ubiquitinylates the 119th lysine residue of histone H2A, strengthening gene transcription repression. PCGF and PHC confer structural scaffold support to PRC1. Reproduced from [97]. Figure made with Biorender.com

PART II: Impact of DNA replication origin positioning on transcription

1.3. Transcription and transcription-related elements

Many studies have looked at the impact of transcription and transcription-related epigenetic marks on the activity, choice and timing of DNA replication initiation [8, 15, 54, 56, 109-113]. However, few studies ever examined the potential impact of DNA replication initiation positioning and activity on gene regulation (discussed below). Does DNA replication initiation impact gene regulation? If so, can DNA replication origins be considered regulatory genomic elements, similar to enhancers or CTCF-binding sites? [114]. In this part of the thesis, I began to explore the potential impact of DNA replication initiation and/or activity on transcription.

Licensing of DNA replication origins starts in the G1-phase of the cell cycle and origins are later activated and fire through S-phase [26]. During the DNA replication process, DNA is also used as a template for transcription, where DNA is copied by the RNA polymerase into coding or non-coding RNA (ncRNA), being translative or non-translative RNA, respectively. Due to the co-localization of DNA replication and transcription, the cell likely confers some coordination, and probably some cross-effects upon dysregulation, between the two biological processes [26, 115]. While numerous studies have looked at the potential collision between transcription and replication machinery (and its consequences for genome integrity) [116, 117], few examined DNA replication initiation as a gene regulatory mechanism.

Effective transcription occurs in the cell by the coordinated involvement of transcriptional regulatory elements, which include regulatory nucleotide sequences of a gene plus regulatory protein elements. Regulatory sequence elements include the promoter, plus response elements, such as enhancer sequences where transcription factors bind and promote RNA polymerase binding or silencer sequences to which repressive proteins block RNA polymerase binding to prevent expression of the genes.

1.3.1. Gene regulatory elements

1.3.1.1. Promoters

A promoter region is a sequence of DNA to which transcription factors bind to regulate and promote transcription of the gene downstream of the promoter. Promoters are located upstream near the transcription start sites of genes (TSS) and can be about 100-1000 bp long depending on transcription product, cell type and species. In bacteria, the promoter is recognized by the bacterial RNA polymerase and an associated (sigma) factor. Moreover, bacterial promoters are highly conservated in time and are defined by two essential elements that are found in most bacterial promoters: the -10 element (Pribnow Box, TATAAT) and the -35 element, (TTGACA), being - upstream of the TSS. However, eukaryotic promoters are more complex and diverse, making them difficult to characterize [118]. Eukaryotic promoters typically contain the transcription start site (TSS), a binding site for RNA polymerase, and several regulatory elements, which can be found at several kilobases away from the TSS (enhancers or transcription factor binding sites) [119]. The transcriptional complex causes the DNA molecule to bend back on itself, allowing the regulatory elements to get closer to the TSS. The eukaryotic promoters can contain consensus sequences called TATA box (TATAAA, present in 24% of promoters) and B recognition element (BRE, present in 22% of promoters), located close to the TSS (30-40 bp upstream) and recognized by the ubiquitous transcription factors TATA-binding protein (TBP) and TFIIB, respectively [119-122]. These regulatory sequences typically bind at least 7 different proteins called transcription factors, which are involved in the formation of the transcriptional complex. All these elements make promoters critical elements in the process of regulation of the transcription levels.

A conserved element found in numerous vertebrate promoters near the TSS (~250 bp upstream, called proximal promoters) is CpG islands (CGIs) [56, 123]. It is currently believed that CpG islands can serve as TSSs. CGI-containing promoters typically lack TATA boxes and have more dispersed transcription initiation [56]. Cytosines in CpG dinucleotides can be methylated (5-methylcytosines) by DNA methyltransferases. Methylated CpG cytosines confer 70-80% of the total CpG cytosines in mammals [124]. The presence of multiple methylated CGIs of promoters causes stable silencing of genes [125]. CGIs are also frequently present in promoters for functional non-coding RNAs (ncRNAs) such as microRNAs [26, 126].

1.3.1.2. Enhancers

Among the assemblage of elements, enhancer sequences and the transcription factors recruited by enhancers have a major role in positive gene expression [127-129]. Enhancers are short (50-1500 bp) cis-acting regions that control cell-type-specific gene expression programs by looping long-distance (up to thousands of nucleotides) DNA regions to come proximal to promoters [130].

1.4. DNA replication initiation and transcription

Recent global studies in vertebrates have reported that the positioning of DNA replication origins in the genome is not random; instead, DNA replication origins are preferentially positioned near promoters [8, 16, 18, 43, 50, 62], usually in the proximity of TSS, a feature that could partly explain the association of origins with CpG islands (also found in promoter regions) [56]. Interestingly, origins found near promoter regions are more concentrated in number and have higher activity than non-promoter-associated origins [8, 15, 131]. To date, it is not known whether the specific localization of DNA replication initiation to promoters has a functional consequence on transcription.

1.4.1. Replication timing and regulation of transcription

DNA replication origin density of a given locus tightly correlates with the timing of the replication of that loci; such that origin-dense regions (mostly found in open chromatin) are replicated first during S-phase [111] and origin-poor areas (found in heterochromatin) are replicated last [8, 21, 27, 42, 47, 54, 70, 132]. This feature ensures that the genomic regions important for the cell are replicated first and are packaged into what is called compartment A (active compartment of the nucleus that contains euchromatin) [26]. Despite these findings, which suggests that there is the existence of coordination between replication and transcription as well as coexistence, it is still unclear whether the positioning of transcriptional promoters helps establish active DNA replication origins and/or vice versa [26, 115].

As reviewed by N. Rhind and D.M. Gilbert [54], replication positioning and timing correlate with gene expression in a range of species (from humans to the fruit fly). For example, early replication was found to have a strong correlation with transcriptional activity in multicellular organisms during developmental differentiation [109, 132, 133]. Causal mechanisms for this correlation have not yet been described. Unlike multicellular organisms, no such correlation

can be found in the model species budding yeast, except for a few regions of heterochromatic regions (telomeres), which are replicated late in S-phase [134, 135].

In an experiment performed by the C. Nieduszynski Lab [136], replication timing was artificially delayed for the histone HTA1-HTB1 gene pair in *S. cerevisiae*, which in turn resulted in a ½-fold decrease in the expression of histone genes. Of note, as reported for *S. cerevisiae* and other species, histone gene pairs are closely positioned to replication origins [137] and are expressed during S-phase [138-140].

Taken together, these studies give hints for a potential influence of DNA replication positioning and timing on nearby gene expression.

1.4.2. Implications of replication origin dysregulation on transcription

The importance of correct origin licensing regulation is emphasized by the consequences of its dysregulation. Replication origin dysregulation can lead to notorious consequences in the cell such as DNA damage and genome instability, and such disturbances can have deleterious phenotypes and even compromise the viability of the cell [26, 115]. As recently reviewed by M. Saponaro, one of the causes of deleterious phenotypes is impairment of gene expression [115]. As an example, L. Williamson et al. reported that DNA damage (UV irradiation) can affect the gene expression of nearby promoters (within 25 Kb) in human cells [141]. When certain oncogenes are overexpressed, such as cyclin E (G1-phase cyclin), a premature entering of DNA replication initiation can occur. Premature S-phase entering results in DNA replication stress intermediates [11] and improper activation of a set of origins that are not activated in healthy cellular conditions [142, 143]. These set of origins are positioned in highly transcribed regions, which, under replication stress situations can generate conflicts between replication and transcription, often leading to fork collapse and erroneous transcription products [143]. Conflicts between the transcription and replication machineries, such as the collision between RNA polymerases and DNA helicase, have been observed *in vitro* and *in vivo* from bacteria to eukaryotic probes [116, 117].

These observations indicate that origin dysregulation and activation of DNA damage repair pathways can impact local transcription and create stressful situations in the DNA in the replication-transcription interface [115]. Despite these grave consequences, mammalian DNA replication origins tend to localize to promoter regions, where the chances of collisions are higher. I thus wished to examine, if DNA replication positioning at promoters has a functional consequence for transcription, in an unperturbed cell cycle.

1.4.3. Genes transcriptionally investigated

Few studies examined the direct impact of deletion of an origin on nearby gene activity in mammals. P. Prorok et al. generated a cell line in which a DNA replication origin was inactivated [41]. In this study, the origin activity was significantly reduced following the removal of a G4 sequence of an endogenous origin (Ori1) in both alleles using Crispr-Cas9 technology (Figure 4, Supplementary Figure 2). The G4 element was located close to the promoter of the Rai1 gene [144]. However, P. Prorok et al. [41], reported no impact on Rai1 expression in asynchronously growing cells. The study did not check the expression of the flanking genes, named Pemt (upstream) and Srebf-1 (downstream), as shown in Figure 4. In this thesis, I thoroughly examine the impact of DNA replication origin deletion on the local transcription landscape using sensitive assays and the cell line generated by P. Prorok et al.

Of note, the researchers did not account for the expression timing of the Rai1 gene through the cell cycle, a relevant feature considering the association of the gene's expression with circadian clock regulation.

1.4.3.1. Retinoic acid-induced 1 (Rai1) gene

The Retinoic acid-induced 1 (Rai1) gene is a dosage-sensitive, highly conserved gene among species, with functions such as embryonic neurodevelopment, cell cycle and growth regulation [145]. Rai1 gene is found in chromosome 11 in mice and the gene protein product (RAI1) functions as a chromatin-binding protein and transcriptional factor. Duplications of the gene have been associated with Potocki-Lupski syndrome (PTLS), a cognitive and behavioral condition [146, 147]. Moreover, haploinsufficiency of this gene is associated with the Smith-Magenis syndrome (SMS) [148, 149], a developmental disorder that affects intellectual and behavioral abilities. Another major feature of the SMS is the sleep disturbance of the patients. Phenotypic [150] and molecular evidence [151] show that haploinsufficiency and duplications of Rai1 disrupt the circadian rhythm (sleep-wake cycle) through the dysregulation of the cell circadian regulator circadian locomotor output cycles kaput (CLOCK), which causes sleep deficiencies in SMS [146, 151]. Overall, the available data on RAI1 function demonstrate that dysregulations in the expression of the gene impact the circadian machinery gene expression [146]. Thus, the impact of expression of the Rai1 gene by the deletion of the endogenous origin Ori1 could potentially cause phenotypes such as Smith-Magenis syndrome.

1.4.3.2. Phosphatidylethanolamine-N-methyltransferase (Pemt) gene

The Pemt gene is found upstream of Rai1 on chromosome 11 (mouse) (Figure 4). Pemt encodes for a phosphatidylethanolamine N methyltransferase enzyme, which catalyses the conversion of phosphatidylethanolamine to phosphatidylcholine in the liver (Figure 1 of [152]). This reaction is crucial for liver homeostasis, cellular membranes integrity, and lipid extracellular metabolism and intracellular signaling [153]. However, Pemt expression is not limited to the liver. A study performed by L.D. Gorné et al. reported a prominent Pemt expression variation during the day, with the highest levels during the night [154]. Further studies are needed to clarify the circadian rhythm regulation of Pemt expression.

1.4.3.3. Sterol regulatory element-binding transcription factor 1 (Srebf-1) gene

The Srebf-1 gene is found downstream of Rai1 on chromosome 11 (mouse) (Figure 4) with its TSS located near the transcription end site of Rai1. Srebf-1 encodes for the sterol regulatory element-binding transcription factor 1 [155], a factor that plays key role in the regulation of the lipid and sterol metabolism by the liver [156]. A study performed by K. Lech et al. has shown that Srebf-1 expression varies during the circadian rhythm in peripheral blood cells [157]. Moreover, Srebf-1 has also been associated with genomic deletions that cause Smith-Magenis syndrome [158].

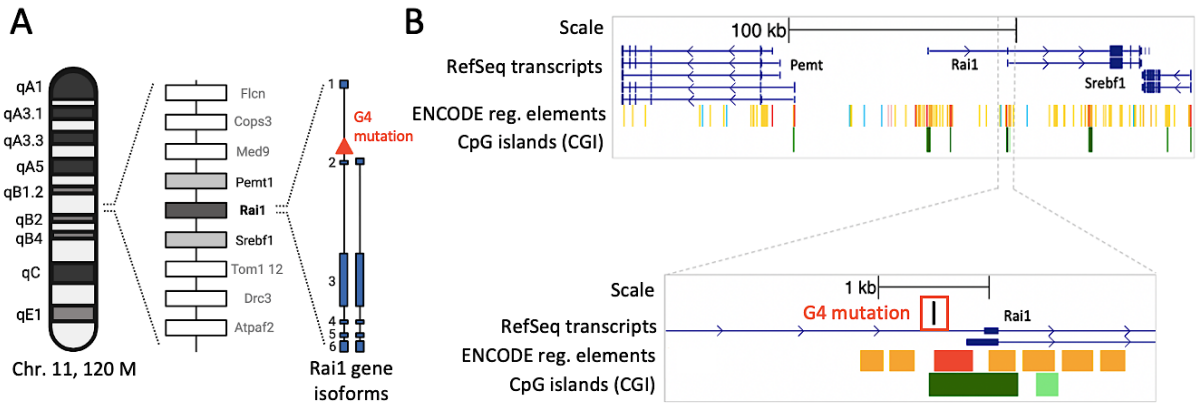


Figure 4. Approach followed in this thesis to test the impact of local gene expression upon deletion of a G4 element. (A) Genom c ocat on of genes w th n the mouse chromosome 11, nc ud ng Ra 1, Pemt and Srebf-1. The ustrat on nc udes the ocat on of the G4 e ement genom c mutat on, performed by P. Prorok et a . [41]. Adapted to mouse from [148]. F gure made w th B orender.com **(B)** UCSC snapshot show ng the s ze and ocat on of Ra 1, Pemt and Srebf-1 genes w th n the mouse genome on chromosome 11. The G4 e ement genom c mutat on s a so marked n red. Co or code: transcr pts (b ue), promoter (red), CpG s ands (green), regu atory e ements (ye ow). The arrows n the transcr pts show the transcr pt on d rect ona ty.

2. HYPOTHESIS, OBJECTIVES & BRIEF EXPERIMENTAL PLAN

2.1. PART I

The positioning of DNA replication origins throughout the genome is not random. Previous studies suggested that Polycomb repressive complexes (PRC) -bound, repressed gene promoters frequently host DNA replication initiation [8, 16, 49]. Thus, I hypothesize that PRC may actively recruit DNA replication initiation activity (origins) to transcriptionally silenced gene promoters.

To test this hypothesis, a mESC cell line containing a tamoxifen-inducible genomic deletion for the catalytic domain of the EZH2 (catalytic subunit of PRC2) [159] was used to perform a genome-wide evaluation of the transcriptome (RNA-Seq), the epigenetic mark landscape of H3K27me3, H3K27ac and H3K4me3 (ChIP-Seq), the accessible chromatin (ATAC-Seq), and the mapped active origins (SNS-Seq).

Here, I aim to analyze the genome-wide impact of loss of PRC2 activity on DNA replication initiation activity in mammalian cells and whether this is mediated by transcription or chromatin landscape.

2.2. PART II

Given the evidence from model organisms [109, 136, 160], I hypothesize that the positioning of DNA replication initiation sites (origins) may directly impact/regulate the amount of transcription (gene expression) in mammalian cells.

To test this hypothesis, a NIH 3T3 cell line containing an irreversible genomic deletion for a G-quadruplex (G4) element, which deletion was previously associated with loss-of-function of a single DNA replication origin (Ori1) located near the Rai1 gene promoter, and its matching wild-type line were used [148]. The gene expression of Rai1, Pemt and Srebf-1 genes (neighboring upstream and downstream genes, respectively) has been previously associated with the rhythmicity of the circadian clock (which is in turn regulated by the cell cycle) [146, 151, 152, 154, 161]. Due to this reason, the expression of these genes will be assessed using Q-PCR from synchronized cells at different time points of the cell cycle.

I aim to study the direct impact of loss-of-function of a single origin (G4 element of Ori1) on local transcription (gene expression of Rai1, Pemt and Srebf-1) in mammalian cells.

3. MATERIALS & METHODS

3.1. Tissue culture

3.1.1. Cell sources and culture

The tamoxifen-inducible Ezh2$^{\Delta SET}$ mouse embryonic stem cell (mESC) line (Supplementary Figure 1A) was a gift from Amanda Fisher (C.F. Pereira et al., Cell Stem Cell. 2010) [159]. These cells were cultured in Prof. Marcel Méchali's Lab according to their standard protocols. In brief, cells were cultured on gelatin-coated dishes in Dulbecco's Modified Eagle's Medium (DMEM) containing 15% fetal bovine serum (FBS, HyClone), 2 mM L-glutamine (Gibco), 1% non-essential amino acids (MEM NEAA, Gibco), 1% nucleoside mix (Gibco), 40 mM b-mercaptoethanol (Gibco), and 1,000 U/mL recombinant Leukemia Inhibitory Factor (LIF, Milipore). For excision of the SET domain, cells were incubated with 100 nM Tamoxifen (Sigma-Aldrich) for 48h and grown for the next 48h in cell culture medium without Tamoxifen (Tmx) supplementation.

The mouse embryonic fibroblast cell line NIH 3T3 (ATCC CRL-1658) in which a DNA replication Origin G-rich Repeated Element (OGRE/G4) was knocked out by Crispr-Cas9 and its matching wild-type line were a gift from Prof. Marcel Méchali Lab (Prorok et al., Nat. Com., 2019) [41] (Supplementary Figure 2A). The clone Dr. I. Akerman Lab received contained 2 modified alleles, as shown in Supplementary Figure 2B-C. NIH 3T3 cells (3T3^{WT} and 3T3$^{\Delta G4}$) were cultured in high glucose DMEM (Sigma-Aldrich, D5796) culture media supplemented with 10% FBS (Gibco, Thermo Fisher Scientific), and 50 U/mL Penicillin-Streptomycin (Gibco, Thermo Fisher Scientific), in a humidified incubator with 5% CO_2 at 37ºC. Modified NIH 3T3 cells were passaged as follows: Once cells reached 80-100% confluency, the culture media was removed, and cells were rinsed once with 10 mL warm 1X Dulbecco's phosphate-buffered saline (D-PBS) (Gibco, Thermo Fisher Scientific). 1.5 mL/100 mm dish and 0.05% trypsin-0.53 mM EDTA 4Na (Gibco, Thermo Fisher Scientific) were added to the surface-attached cells, and the flask was placed in the incubator for 1-2 min. Detachment was confirmed under an inverted microscope. Trypsinization was stopped with complete media, and cells were separated by slowly pipetting up and down. The trypsinized cells were centrifuged at 1000 x g for 4', the supernatant containing trypsin was removed, and the cells were resuspended in trypsin-free complete media and plated in a total volume of 10 mL for 100 mm dishes, and 2 mL for 6-well plates. Dilutions from 1:5 to 1:15 were used.

All cell culture procedures were performed under aseptic conditions inside a biohood and sterile disposable material. Biosafety level 2 (BMBL, ATCC).

3.1.2. Cryopreservation and thawing

When necessary, cells were cryopreserved in liquid nitrogen (-196ºC) in 90% FBS and 10% DMSO BioUltra (Sigma-Aldrich). Vials were thawed by gently swirling the cryovials in a 37ºC water bath and then diluted with 15 mL pre-warmed DMEM culture media in T75 flasks (Sigma-Aldrich). Cells were left growing in the incubator for 24 hours and the media was replaced with fresh media. When cells were brought back from cryopreservation, at least two passages (> one full week) were allowed to enable cell recovery, resuming consistent cell cycle progression before starting the experiments.

3.2. Cell synchronization and sorting

The University of Birmingham Biomedical Sciences BSc student Sam Brain contributed to this work by performing some of the replicates of synchronization experiments under my supervision.

3.2.1. Cell synchronization and release from serum starvation

This method was used to synchronize both $3T3^{WT}$ and $3T3^{\Delta G4}$ into G0/G1 which were released back into the cell cycle in a synchronous fashion. Fully confluent cells were seeded at 20% (1/5) confluence into 100 mm dishes and grown for 24 hours twice in complete DMEM media (10% FBS) before media replacement with 12 mL of DMEM serum starvation media (0.5% FBS) and maintained in the incubator for another 24h. At 24h, cells were synchronized in G0. A 48h synchronized sample was also obtained by culturing cells for 48 hours in starvation media (here referred as G0-48h). To release the cells from 24h serum starvation, the serum starvation media was removed, cells were washed with warm 1X D-PBS, and then split 1:3 ratio returning to complete DMEM media (10% FBS). Samples for FACS analysis and RNA extraction were harvested at 0h (G0), 2h (G0/G1-phase), 4h (G1-phase), 6h (G1-phase), 8h (G1/S-phase), 16h (S-phase), and 18h (S/G2-phase) post-release. In addition, samples were harvested from asynchronously (AS) growing cells as well as cells with long serum starvation (G0-48h, 48 hours).

3.2.2. Sample preparation, staining and fluorescence-activated cell sorting

Samples for FACS analysis were collected as follows; attached cells were trypsinized as described above and the supernatant was removed. The cell pellets were resuspended with 3 mL of ice-cold 70% ethanol and stored at 4ºC, until use (< 1 week).

Once all samples were collected, the 15 mL tubes were centrifuged at 1000 x g for 4 min and the supernatant was removed from the fixed cells. The pellet was then resuspended in 300 µL of PBS containing 0.01 µg/mL RNase A (Sigma-Aldrich) and 0.5 µM Vybrant DyeCycle Violet DNA staining (Thermo Fisher Scientific). Samples were incubated for 1h in a 37ºC water bath in the dark.

Cells were sorted using a CytoFLEX Flow Cytometer (Beckman Coulter Inc.). An unstained sample was used to localize the cells and optimize flow. The laser 405 (Violet), with a bandpass filter of 450/50 PB450 was used to measure the signal intensity of Vybrant DyeCycle Violet Stain (369/437). Once the parameters were identified, three gates were set to specifically select the % cells counted by DNA content: G1, S and G2/M. Sorting was performed at 15-30ºC from 5 mL FACS tubes (Thermo Fisher Scientific). 10,000 events were recorded at slow flow rate (10 µL/min) for each sample.

3.3. Local gene expression evaluation

3.3.1. RNA extraction for quantitative PCR

To isolate RNA from cell culture, DMEM was removed and samples were washed with D-PBS. Following this, RNA extraction was performed with the RNeasy Mini Kit (QIAGEN), following manufacturer's instructions. After washing with D-PBS, the RNeasy buffer (300 µL/6-well) was added to each sample, placed in an Eppendorf tube, and stored in a -80ºC freezer until RNA extraction. The purified RNA samples (25-30 µL) were stored at -80ºC until cDNA synthesis.

3.3.2. cDNA synthesis for quantitative PCR (reverse transcription)

Before reverse transcription, the concentration of the RNA samples was measured with NanoDrop™ Spectrophotometer ND-1000 (Thermo Fisher Scientific) at 260/280 nm from 1.5 µL drop RNA samples.

The reverse transcription (RT) of RNA to cDNA was performed using the Invitrogen SuperScript IV Reverse Transcriptase Kit (Thermo Fisher Scientific) on a ThermoCycler machine, following manufacturer's instructions. The 10 µL reaction was carried out in two steps in single tubes of an Xtra-Clear PCR 8-Strip (StarLab): a first annealing step mix of cDNA, random hexamers and dNTP Mix (Invitrogen, Thermo Fisher Scientific), and a second RT reaction step of the previous mix and SSIV buffer (SuperScript IV Kit), DTT (SuperScript IV Kit), RNaseOUT™ Recombinant Ribonuclease Inhibitor (Invitrogen, Thermo Fisher Scientific), and the SuperScript IV RT enzyme. The first step was incubated at 65ºC for 5' and the second step was incubated at 23ºC for 10', at 50-55ºC for 10', and at 80ºC for 10'. The RT reactions (cDNA) were stored at -20ºC until Q-PCR performance.

3.3.3. Primer design and quality controls

Primers were designed with PrimerBank (MGH-PGA) considering G/C contents of 40-60%, amplicon sizes of 90-150 bp, and melting temperatures (Tm) ~ 60ºC. Moreover, the candidate primers were checked with UCSC In-Silico PCR and PrimerBLAST NCBI upon the mouse genome Dec. 2011 (GRCm38/mm10). Primers were designed against the Enhancer-of-Zeste homolog 2 gene, mEzh2 (Gene ID 14056); Retinoic acid-induced 1 gene, mRai1, (Gene ID 19377); Sterol Regulatory element-binding transcription factor 1 gene, mSrebf-1 (Gene ID 20787) and Phospatidylethanolamine-N-methyltransferase gene, mPemt (Gene ID 18618). A primer was designed to detect the Ezh2-SET domain deletion upon Tamoxifen recombination (Supplementary Figure 1A). Primers were designed to detect all isoforms from mRai1 gene (in exon 3, Supplementary Figure 2B) and the long isoform of mRai1 (in exon 1). Primers against mTbp (Gene ID 21374) and mActb (Gene ID 11461) were also used to normalize the expression of the other genes.

Primers were purchased in Sigma-Aldrich and were resuspended upon arrival with DNase/RNase-Free distilled water to reach a concentration of 100 µM. Before experimentation, primer performance was tested with control cDNA and negative controls.

Primer sequences can be found in Table 1.

Gene	Use	Oligo sequence (5' to 3')		Amplicon size (bp)
		Forward primer	Reverse primer	
mEzh2	Detect ng the Ezh2-SET doma n de et on	CTGCTCTGAATGGCAACTCC	TTATTCATAGAGCCACCTGG	436
mRa 1	Dectect ng a soforms	AGGAACACTGGGTCCATGAG	GCACATGGGTAGTGGTAGGT	190
mRa 1	Detect ng ong soform	TGTCGAGGTCCGAGGAAG	CCCCTCGCTTCTTCCAGTA	185
mPemt	Dectect ng a soforms	ATCACCATTGTGTTCAACCCA C	ATCACCATTGTGTTCAACCCA C	112
mSrebf-1	Dectect ng a soforms	GATGTGCGAACTGGACACAG	CATAGGGGGCGTCAAACAG	104
mTbp	Load ng contro	ACCCTTCACCAATGACTCCTA TG	ATGATGACTGCAGCAAATCG C	190
mActb	Load ng contro	GGCTGTATTCCCCTCCATCG	CCAGTTGGTAACAATGCCAT GT	154

Table 1. Primer sequences designed with supporting information. Tab e show ng the sequences for the forward and reverse pr mers and the s ze of the reg ons that the pr mers amp fy (amp cons). The pr mers are des gned for the genes transcr pt ona y nvest gated n th s thes s (Ezh2, Ra 1, Pemt and Srebf-1). Tbp and Actb are the oad ng contro pr mers.

3.3.4. Quantitative PCR

Quantitative real-time PCR was performed using SYBR^{TM} Green PCR Master Mix (Thermo Fisher Scientific) in a Real-Time PCR ABI 7900HT 384-Well Block Module machine (Thermo Fisher Scientific) with a final reaction volume of 10 µL and a final primer concentration of 250 nM. The cDNA samples were diluted by adding 40-50 µL of distilled water (depending on previous NanoDrop^{TM} measurements) and 4 µL of sample was used in each reaction along with 6 µL of reaction mix. Each sample was tested in triplicates. Distilled water was used as a negative control. A standard curve was used to quantify the samples. The standard curve was made with a 10-fold serial dilution from a mixture of concentrated cDNA from asynchronously growing 3T3^{WT} and 3T3$^{\Delta G4}$ cells. Gene expression was normalized to mTbp expression.

The following assays were performed by Dr. Paulina Prorok or Lab technicians in the Lab of Prof. Marcel Méchali. The data generated was then analyzed by myself (MSc candidate Nerea Murugarren) with the support of my supervisor, Dr. Ildem Akerman. The protocols are provided below as they directly relate to the analysis of the data.

3.4. Sequencing

3.4.1. RNA extraction, and sequencing

Approximately 5 million mESCs from n = 3 independent cell cultures each between $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ were collected for total RNA extraction using the RNeasy Mini Kit (Qiagen). An RNA integrity number (RIN) was calculated for every sample, being RIN >8 an acceptable level of RNA degradation for sequencing. Contamination ratios (obtained through Q-PCR) also helped assess the adequate sample purities. Libraries were prepared using the Illumina TruSeq Stranded mRNA Sample Preparation Kit and sequenced using an Illumina HiSeq 2500 apparatus at the Montpellier GenomiX (MGX) facility (50 bp, single-end reads).

3.4.2. RNA-primed short nascent strand (SNS) isolation, and sequencing

SNSs were purified as previously described in detail [49, 50]. Briefly, cells were lysed with DNAzol (Invitrogen), then nascent strands (NS) were first separated from genomic DNA based on sucrose gradient size fractionation. Fractions corresponding to 0.5-2 Kb were isolated and phosphorylated by T4 polynucleotide kinase (Thermo Fisher Scientific) and digested with two consecutive overnight rounds of λ-Exonuclease (λ-Exo, 100-150 units, Thermo Fisher Scientific) to eliminate contaminating DNA. A third λ-Exo digestion (100 units) was performed for 2 h at 37°C. Three samples were purified from n = 3 independent cell cultures. The RNase A controls were obtained by treating NS-containing fractions with 50–100 µg/mL RNase A (Sigma-Aldrich) before λ−Exo digestion to remove RNA primers at the 5′ ends of the NSs. The quality of origin enrichment in each sample was determined by Q-PCR amplification using the specific primers listed in Table 1 with the LightCycler 480 SYBR Green Master Mix (Roche), in a LightCycler 480 II apparatus (Roche). The NS enrichment was calculated as the ratio of the signal scored at origin-specific and background regions.

Single-stranded NS DNA and RNase A treated control DNA were first purified using an Illustra™ CyScribe GFXTM purification kit (GE Healthcare), then converted into double-stranded DNA by random priming using DNA polymerase I (Klenow fragment) and the Array

CGH kit (BioPrime). The Illumina TruSeq ChIP Sample Prep Kit was used for the preparation of sequencing libraries following manufacturer's instructions. Samples were sequenced using the Illumina HiSeq 2500 (50 bp, single-end reads) at the Montpellier GenomiX (MGX) facility (see https://www.mgx.cnrs.fr/ for more details).

3.4.3. Chromatin extraction and immunoprecipitation (ChIP), and sequencing

This was carried out according to the protocol by Luco, R. F. et al. [162]. Briefly, samples were crosslinked for 10 min in 1% formaldehyde at room temperature, then incubated for 5 min at room temperature with 125 mM glycine. Chromatin extraction was performed with approximately 1 x 10^7 ESCs from n = 2 independent cell cultures, harvested by centrifugation and washed twice in cold PBS. Pellets were resuspended by rotating in 1 mL of cold lysis Buffer A (50 mM HEPES-KOH pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100) freshly supplemented with protease inhibitor cocktail. The lysate was centrifuged at 2500 rpm for 5 min, and the resulting supernatant was discarded. 1 mL cold lysis Buffer B (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, with protease inhibitors) was added to the pellet, which was rotated at room temperature for 10 min, then centrifuged at 2500 rpm for 5 min, and the supernatant discarded. Extraction of chromatin was performed by adding 750 µL of Buffer C (10 mM Tris-HCl pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% sodium deoxycholate, 0.5% N-lauroylsarcosine, with protease inhibitors) to the remaining pellets. Then, purified chromatin was sonicated in an ultrasonic bath (Diagenode Bioruptor) to generate DNA fragments of average length of 200–500 bp, and insoluble material was pelleted. Sonicated chromatin was immunoprecipitated overnight with 1 µg antibody in a final volume of 500 µL ChIP buffer (20mM Tris-HCl pH8, 2mM EDTA, 150mM NaCl, 0,1% SDS, 1% Triton, with protease inhibitors). Immunocomplexes were recovered with 30 µL of G-agarose beads (Dynabeads™ Protein G for Immunoprecipitation 10004D; Thermo Fisher Scientific) slurry incubated for 2 h at 4°C. Beads were washed and eluted in 100 mL of elution buffer (1% SDS, 10 mM EDTA, 50 mM Tris-HCl pH 8) for 25 min at 65°C after vigorous vortexing. DNA complexes were de-crosslinked at 65°C overnight, treated with proteinase K and purified with QIAquick PCR purification columns (Qiagen), according to the manufacturer's instructions. Immunoprecipitated DNA was analyzed using SYBR Green (ABI) in a real-time Q-PCR machine (Roche LightCycler). Primers utilized in ChIP-Q-PCR validations are listed in Table 1. The Illumina TruSeq ChIP Sample Prep Kit was used for the preparation of sequencing libraries. Samples were sequenced using the Illumina HiSeq 2500 (50 bp, single-end reads) at the Montpellier GenomiX (MGX) facility.

3.4.4. Antibodies used in this study

Anti-phosphorylated CHK1 (Cell Signaling, 2341 S, dilution 1/1000), anti-PCNA (Sigma, P8825, dilution 1/2500), anti-CHK1 (Abcam, ab32531 , dilution 1/1000), anti-OCT4 (Abcam, ab19857, dilution 1/1000), anti-EED (Sigma-Aldrich, 09-774, dilution 1/1000), anti-EZH1 (Abcam, ab28988, dilution 1/1000), anti-EZH2 (Abcam, 49-1043, dilution 1/1000), anti-β-ACTIN (Sigma, A4700, dilution 1/1000), HRP-linked ECL anti-mouse IgG (GE Healthcare, NA931V, dilution 1/4000), HRP-linked ECL anti- rabbit IgG (GE Healthcare, NA934V, dilution 1/4000) antibodies were used. Information regarding antibody validation can be found on the manufacturers' websites.

3.5. Bioinformatics analysis pipeline

All the bioinformatic analysis presented in this thesis was performed by myself (MSc candidate Nerea Murugarren) with the support of my supervisor, Dr. Ildem Akerman.

The bioinformatics analysis on this project was performed with the use of the University of Birmingham supercomputer HPC and HTC cluster BlueBEAR service (RedHat 8 OS), which provides *a High-Performance Computing* service to the University's research community (see *http://www.birmingham.ac.uk/bear* for more details). The UNIX shell scripting language "bash" was used to interact with the server and perform the following analysis. Each of the following procedures was performed using a bash or R script or the Linux command line. The bioinformatics pipeline is represented in Figure 5 as an illustration.

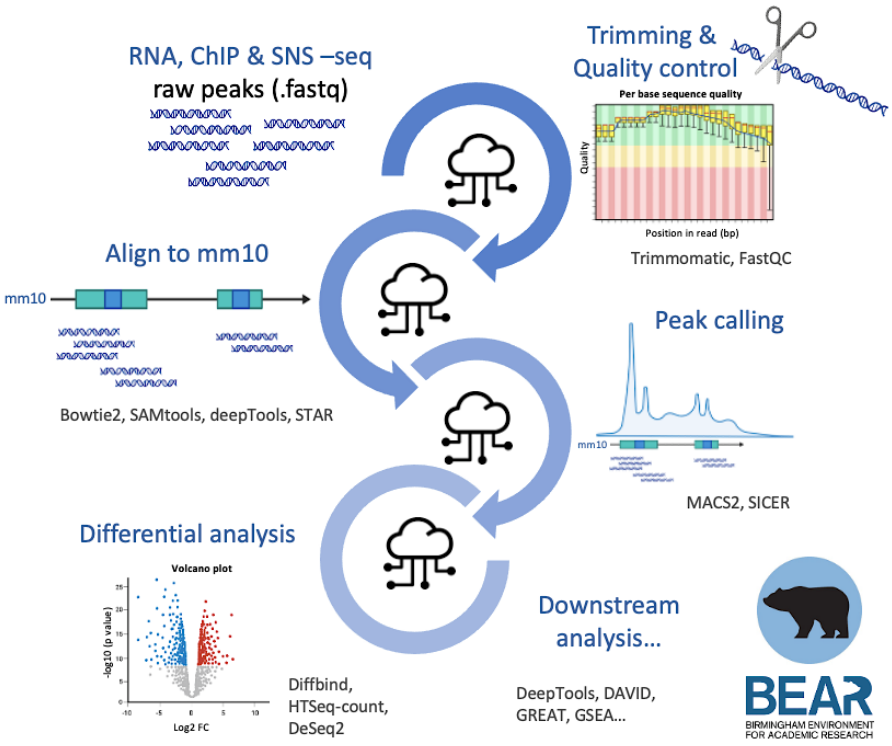


Figure 5. Overview of the bioinformatics pipeline followed in this thesis. The f gure shows an overv ew of the b o nformat cs p pe ne and the too s used to ana yze the sequenc ng data presented n th s thes s. The b o nformat cs ana ys s of th s thes s has been performed w th the use of the Un vers ty of B rm ngham supercomputer HPC and HTC c uster B ueBEAR serv ce through "bash" scr pt ng anguage.

3.5.1. Data accession

Sequencing data in this study was generated by the Montpellier GenomiX (MGX) facility with an Illumina HiSeq 2500 apparatus using the TrueSeq ChIP Library Preparation Kit (Illumina) (between $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ RNA-Seq, H3K27me3 and H3K27ac ChIP-Seq, and SNS-Seq). Original raw .fastq data was downloaded using FTP and files were checked for integrity using MD5sums. Raw and processed data will be made available on GEO [163].

Moreover, publicly available data were downloaded and analyzed with my pipeline (explained as follows) to support the sequencing data obtained in Prof. M. Méchali. To download the .fastq files from the GEO database, the bash prefetch and fasterq-dump commands (SRA-Toolkit [164] v2.9.6.1) were used with each SRR file of interest. The downloaded files were verified using vdb-validate (SRA-Toolkit v2.9.6.1) (see https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software for more info). Downloaded data sources can be found in Table 2.

Dataset	Cell line	Source [Ref]	GEO
EZH2 b nd ng s tes (ChIP-Seq)	E14	PMID: 24141703 [165]	GSE49431
EZH2 b nd ng s tes (ChIP-Seq)	mESC, 129Sv/J	PMID: 29932419 [166]	GSE94834
Chromat n access b ty (ATAC-Seq)	LSK, C57BL/6	PMID: 31189531 [167]	GSE132077
H3K4me3 andscape (ChIP-Seq)	LSK, C57BL/6	PMID: 34732703 [168]	GSE132078
H3K27me3 andscape (ChIP-Seq)	LSK, C57BL/6	PMID: 34732703 [168]	GSE132078

Table 2. Sources for the publicly available data utilized in this thesis. Tab e show ng the sequenc ng dataset, the ce ne where the sequenc ng data was obta ned from, and pub c y ava ab e sources. The pub c ava ab e data was retr eved from the GEO database. Pub c y ava ab e data has been obta ned for EZH2 b nd ng s tes, chromat n access b ty and H3K4me3 & H3K27me3 marks andscape.

3.5.2. Alignment and quality control of raw sequencing data

Fastq files were produced with a bcl2fastq version 2.17 from Illumina. The quality of the produced reads (50 bp, single-end) from each ChIP-Seq, SNS-Seq and RNA-Seq sample was assessed before and after trimming with FastQC [169] (v0.11.9). Trimming was performed with Trimmomatic [170] (v0.39) (parameters 3:10:8:6 LEADING:5 TRAILING:5 SLIDINGWINDOW:4:15 HEADCROP:3 MINLEN:27). Prior to alignment, the latest version of the mouse genome (GRCm38/mm10, UCSC, Dec. 2011) was downloaded and indexed using the STAR genomeGenerate command [171]. The STAR indexed splice junction sites were suitable for the 50 bp (--sjdbOverhang 49), single-ended reads. For ChIP-Seq and SNS-Seq, the reads were then aligned to the mouse genome (GRCm38/mm10, UCSC, Dec. 2011) using Bowtie2 [172] (v2.3.5.1), and the produced .sam files were converted to .bam and .bw files using SAMtools [173] (v1.10) and deepTools [174] (v3.5.0) (parameters --normalizeUsing CPM --binSize 10), respectively. For ATAC-Seq, obtained as publicly available data (see Table 2). The same bioinformatics pipeline to ChIP-Seq data was applied for ATAC-Seq, with an extra normalization. According to the researchers that performed ATAC-Seq [167], the samples were not normalized due to batch effects. Hence, before converting to .bw files, ATAC-Seq .bam was normalized according to a DeSeq2 [175] (v1.18.0 in R v4.1.0) normalizing parameter from quantified peaks, which were previously obtained with R-package Diffbind [176] (v3.9). For RNA-Seq, the reads were aligned to the mouse genome using STAR [177] (v2.7.2b-GCC-8.3.0) (parameters —runMode alignReads --outFilterMultimapNmax 20 --outReadsUnmapped Fastx_failed --outSAMtype BAM SortedByCoordinate). All the samples had high alignment rates, typically > 80%.

3.5.3. Peak calling

SNS-Seq peaks were called using two peak calling programs MACS2 [178] (v2.2.1) and SICER [179] (v1.1). MACS2 (additional parameters –bw 500 -p 1e-5 -s 60 -m 10 30–gsize 2.7e9) was used to identify narrow peaks of replication initiation, the subsequent peak calling using SICER (parameters: redundancy threshold = 1, window size (bp) = 200, fragment size = 150, effective genome fraction = 0.85, gap size (bp) = 600 and FDR = 1e-3) permitted to identified broader zones of replication initiation. MACS2 narrow peaks intersecting with SICER peaks (using BEDTools [180] intersect v2.30.0) were considered as sites of DNA replication initiation. ChIP-Seq peaks (H3K27me3, H3K4me3, H3K27ac and EZH2 binding sites) and ATAC-Seq peaks were called using MACS2 (v2.2.7.1) (additional parameters --bw 500 -p 1e-5 -s 60 -m 10 30 --gsize 2.3e9) in obedience with the ENCODE standards [181] (broad peaks

for H3K27me3 and narrow for H3K27ac). The generated MACS2 peaks were merged and intersected (parameter -wa) with BEDTools [180] (v2.30.0) and intersected to obtain bed files of all peaks, wild-type peaks and knockout peaks (H3K27me3 and H3K27ac).

3.5.4. Quantification and differential peak analysis

Quantification of the SNS/ATAC/ChIP-Seq signals was done based on the p-adjusted value and the Bonferroni method using the R-package Diffbind [176] (v3.9), using the peaks obtained previously. The same package was used to define statistical differential peak activity in SNS-Seq and ChIP-Seq samples (Diffbind (v3.9) in between $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ cells using R Studio (v4.1.0). RNA-Seq .bam files were counted with HTSeq-count [182] (v0.12.4) with the flag intersection-nonempty. Gene annotations were downloaded.

3.5.5. Differential gene expression analysis

RNA-Seq data was subjected to differential gene expression analysis using DeSeq2 [175] (v1.18.0 in R v4.1.0) following DeSeq2 manual.

3.5.6. Gene ontology analysis

The GREAT [183] website (v4.0.4) was used to associate genomic coordinates to genes and obtain the biological processes gene ontology. GREAT calculates statistics by associating genomic regions with genes located in the vicinity (here, within 500 Kb) and applying the gene annotations to the regions.

3.5.7. BEDTools

BEDTools [180] (v2.30.0) was used in the study to sort the data (sort -V -k1,1 -k2,2n), merge the peaks (merge), produce shuffled controls (shuffle -chrom), extend the peaks' size (flank -l -r) and overlap (bedtools intersect -wa) the data.

3.5.8. Data presentation

Volcano and violin plots were generated using ggplot2 [184] (v3.1.0) and EnhancedVolcano [185] (v1.10.0) in R (v4.1.0). Bar plots and pie charts were generated in Microsoft Excel (Excel 2019) using data obtained with BEDTools, R, Diffbind and GREAT. Bar plots and line plots were also generated with Prism GraphPad (v9.0.0) for Mac (GraphPad Software, San Diego, www.graphpad.com) using data obtained from the Q-PCR output. For read pile up plots and heatmaps, matrix were computed with Deeptools [174] (v3.5.0) and Python (v3.8.2). IGV (Integrative Genomics Viewer, MIT, v2.10.x) [186] was used to visualize the sequencing data on mm10 genome regions.

3.5.9. Data sources and availability

Mouse gene names and coordinates (EnsDb.Mmusculus.v79) were obtained using the biomaRt library BiocManager [187] in R. Curated Refseq genes were downloaded from UCSC Table browser [188]. Other publicly available data sources used in this study are given in Table 2. Analysis script for SNS-Seq can be found on GitHub [189] https://github.com/iakerman/SNS-Seq. Any other script is available upon request. All raw data generated in this study will be available on GEO database upon publication [163].

3.6. Statistical analysis

The statistical analysis was performed with the two-tailed, paired T-test to assess statistically significant differences in gene expression levels detected in n = 5 independent experiments with Prism GraphPad (v9.0.0) for Mac (GraphPad Software, San Diego, www.graphpad.com).

If not otherwise specified, statistical significance was set at p-values ≤ 0.05.

4. RESULTS

4.1. PART I

4.1.1. Polycomb-repressed promoters harbour DNA replication origins

Studies from Prof. M. Méchali and Dr. I. Akerman laboratories, and others, suggested that DNA replication initiation is strongly associated with open, active chromatin regions of the genome such as transcription start sites (TSS), enhancers and active histone modification marks [14, 16, 19, 25, 43, 46, 48]. Principally, regions of the genome which are correlated with open chromatin and active epigenetic marks are enriched in replication origins and ORC-bound regions [21]. These regions also tend to replicate early in S-phase [54, 70]. Conversely, regions of the genome with inactive chromatin, gene-poor regions, and epigenetic repressive marks, tend to have low origin content [16, 46-48] and tend to replicate later in S-phase [54, 70]. In accordance, I find that 52% of transcriptionally active genes in mouse ESC cells harbour DNA replication initiation sites (origins) at their promoter regions (Figure 6A). Previous studies suggest that non-transcribing genes can also be associated with DNA replication initiation activity albeit at lower levels [16]. My findings in mouse ESC are consistent with previous findings and suggest that ~33% of transcriptionally silent genes harbour DNA replication origins at their promoters (Figure 6A) [16]. Interestingly, of the silent genes that harbour DNA replication origins at their promoters in mESC cells, the majority (80%) also have CGIs at their promoters (Figure 6B). CGI-containing origins only make up 22% of the repressed genes, suggesting that DNA replication origins are specifically enriched at CGI containing repressed genes, but not non-CGI containing repressed genes. CGI-containing gene promoters are bound and repressed by Polycomb-complex group of proteins through histone 3 tail methylation by EZH2 (i.e. H3K27me3 and me2) (Figure 3) [190]. In accordance, 79% of EZH2-bound silent genes vs 20% of non-EZH2-bound silent genes contain a replication origin (Figure 6C), suggesting a strong correlation between the presence of CGIs, PRC components and DNA replication initiation activity at silent gene promoters.

In brief, my results and others suggest that DNA replication origins are specifically associated with Polycomb- bound repressed genes, rather than all repressed genes.

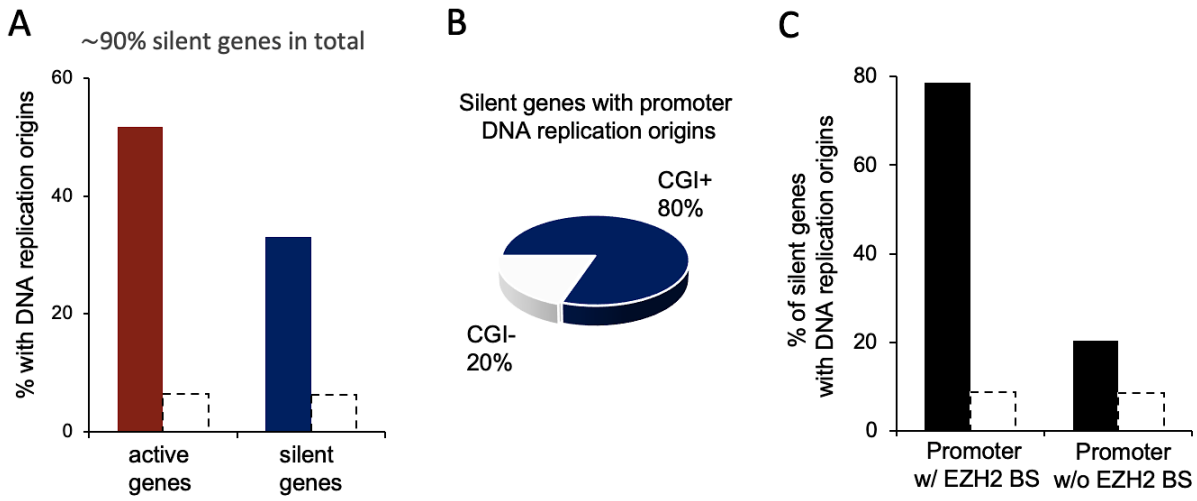


Figure 6. Polycomb-repressed promoters harbour DNA replication origins. (A) Bar p ots represent ng the percentage of over ap between transcr pt ona y act ve genes (act ve genes def ned as hav ng > 3 transcr pts per m on, TPM, n th s ce ne) and transcr pt ona y s ent genes (repressed genes def ned as hav ng < 3 transcr pts per m on, TPM, n th s ce ne) w th DNA rep cat on or g ns (25,357). **(B)** P e chart represent ng the percentage of s ent genes w th DNA rep cat on or g ns at the promoter conta n ng and not conta n ng CpG s ands (CGIs). **(C)** Bar p ots represent ng the percentage of over ap between promoters conta n ng and not conta n ng EZH2 b nd ng s tes (EZH2 BS) w th transcr pt ona y s ent genes conta n ng DNA rep cat on or g ns. The dashed bars show contro over aps w th random zed (shuff ed) reg ons.

I thus asked whether Polycomb-mediated gene repression plays a direct role in the recruitment of DNA replication initiation activity to transcriptionally silent genes in mouse ESC cells. Polycomb-mediated repression was interrupted by a conditional knockout of the catalytic activity of EZH2, an essential component of the Polycomb repressive complex 2, PRC2. We used a published mouse ESC cell line in which the genomic region encoding for the catalytic SET [Su(var)3-9, Enhancer-of-Zeste and Trihorax] [191] domain of Ezh2 can be excised using a tamoxifen-inducible Cre-Lox excision system (C.F. Pereira et al., Stem Cell 2010) [159] (Supplementary Figure 1A). *[Please note Prof. Marcel Méchali's Lab performed the tamoxifen-induced excision of the conditional knockout Ezh2 cell line and all related wet-lab experiments (Supplementary Figure 1B), while I have performed all the bioinformatic analysis presented in this thesis (Figure 5)].* As previously reported [159], tamoxifen-induced excision leads to efficient deletion of exons 14 and 15, generating a premature STOP codon prior to the SET domain coding exons 14 and 15 (Supplementary Figure 1A,C), leading to the degradation of the truncated $EZH2^{\Delta SET}$ ($Ezh2^{\Delta/\Delta}$) protein product without detectable intact EZH2 protein (Supplementary Figure 1D). The truncation of EZH2 did not affect the expression of other components of the PRC complexes (Supplementary Figure 1E). In fact, published work suggests that the PRC2 complex can be assembled in the absence of EZH2 but no histone

marks are deposited onto the chromatin [192]. We also did not observe differences in OCT4 protein level, a pluripotent state marker (Supplementary Figure 1F) between $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ cells suggesting that the $Ezh2^{\Delta/\Delta}$ cells have not differentiated into a specific linage. Notably, normal cell cycle progression in the absence of EZH2 catalytic activity was observed, with similar percentages of cells in active S-phase (Supplementary Figure 1G). This was assessed by the analysis of the cell cycle distribution of $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ cells by flow cytometry pulsed with the thymidine analog BrdU. DNA damage checkpoint activation can result in activation of dormant DNA replication origins, it was thus confirmed that S-phase in $Ezh2^{\Delta/\Delta}$ cells does not result in DNA damage checkpoint activation as assessed by the phosphorylation levels of the checkpoint kinase CHK1 as observed in control cells treated with camptothecin (a topoisomerase 2 inhibitor) at this early time point post-excision (96h) (Supplementary Figure 1H). Taken together, the examination of the $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ cells did not reveal an impact to normal DNA replication progression, checkpoint activation or differentiation state at day 4.

$Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ cells were then subjected to a genome-wide evaluation via sequencing; DNA replication origins were mapped by SNS-Seq, the transcriptional profile was assessed by RNA-Sequencing (RNA-Seq), and the chromatin marks H3K27me3 (repressive mark) and H3K27ac (activating mark) landscape was profiled using chromatin immunoprecipitation followed by sequencing (ChIP-Seq) (Supplementary Figure 1B). I then bioinformatically analyzed both the raw data produced by Prof. M. Méchali Lab in this project as well as publicly available data on EZH2 binding sites, and others, to produce a comprehensive set of genomic data on mouse ESC cells (Methods and Table 2).

4.1.2. Loss of EZH2 catalytic activity results in increased DNA replication initiation

To understand the potential role of EZH2 in DNA replication initiation, Prof. M. Méchali's Lab mapped active DNA replication origin positions using RNA-primed small nascent strand isolation and sequencing (SNS-Seq) in $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ cells. SNS-Seq is a technique that allows the isolation and sequencing of active replication initiation sites primarily by enriching the leading strand start sites, followed by next-generation sequencing (overview of SNS-Seq shown in Figure 1B) [17]. $Ezh2^{+/+}$ cells displayed expected ([23] taken as reference) numbers of SNS-Seq peaks (see Supplementary Table 1), representing sites of frequent DNA replication initiation with peaks occupying both active (mActb) and known repressed PRC-targets (mHox A cluster, [193]) (examples shown in green with an IGV genome browser snapshot in Figure 7A). The Hox cluster is a group of related genes that encode proteins that

specify structural regions of an animal embryo along the head-tail axis of the body [194] and are well established PRC target as further evidenced by the binding of EZH2 (Figure 7A, in orange) [193].

I next wished to examine the differences in SNS-Seq peak activity between $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ cells based on statistical significance using Diffbind. Diffbind is a computational tool that can determine statistically differential accumulation of NGS reads at genomic locations across samples using a TMM (Trimmered Mean of M-values) normalization method, similar to those used for differential gene expression analysis. Examination of differential SNS-Seq signal between $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ cells revealed a predominant upregulation trend with ~5000 peaks showing statistically significant augmentation of activity in $Ezh2^{\Delta/\Delta}$ cells (FDR < 0.05, 8.71% of all origins) (Figure 7B, Supplementary Table 1). Figures 7C-E display read pile up plots showing the signal profile of $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ SNS-Seq samples center on the origins from downregulated, upregulated, and unchanged regions upon loss-of-function of EZH2, respectively. As observed in Figure 7D, upregulated origins reveal a prominent gain of the SNS-Seq signal (origins activity) upon EZH2 loss-of-function. The center of upregulated peaks is surrounded by a “shoulder” (Figure 7D), a shape not found in unchanged peaks (Figure 7E), which suggests that upregulated origins are found in clusters (i.e. origin dense regions). Positive and negative (shuffled) control regions drawn to the same scale do not experience drastic changes in activity, but instead show a flat line (Supplementary Figure 3B, F). The data suggests that loos of EZH2 function leads to a predominant upregulation of activity for a select group of DNA replication origins.

Upregulated origins are enriched in CGIs (60%) and EZH2-bound regions (41%) compared to similarly active control origins, whose activity remains unchanged in $Ezh2^{\Delta/\Delta}$ cells (Figure 7F). Moreover, upregulated origins tend to be located at promoters (64% of Refseq genes, 77% of ENSEMBL genes) (Figure 7G), predominantly at CGI-containing transcriptionally silent promoters (Figure 7H). While previous studies suggested that DNA replication origins are enriched at gene promoters [16, 18], the observed enrichment for upregulated origins at promoters surpasses that of control origins suggesting that upregulation of DNA replication initiation activity specifically takes place in promoter regions (Figure 7G, H).

Taken together, a large fraction of the DNA replication origins that are upregulated upon loss-of-function of EZH2 are located in CGI-containing, repressed gene promoters and overlap EZH2 binding sites.

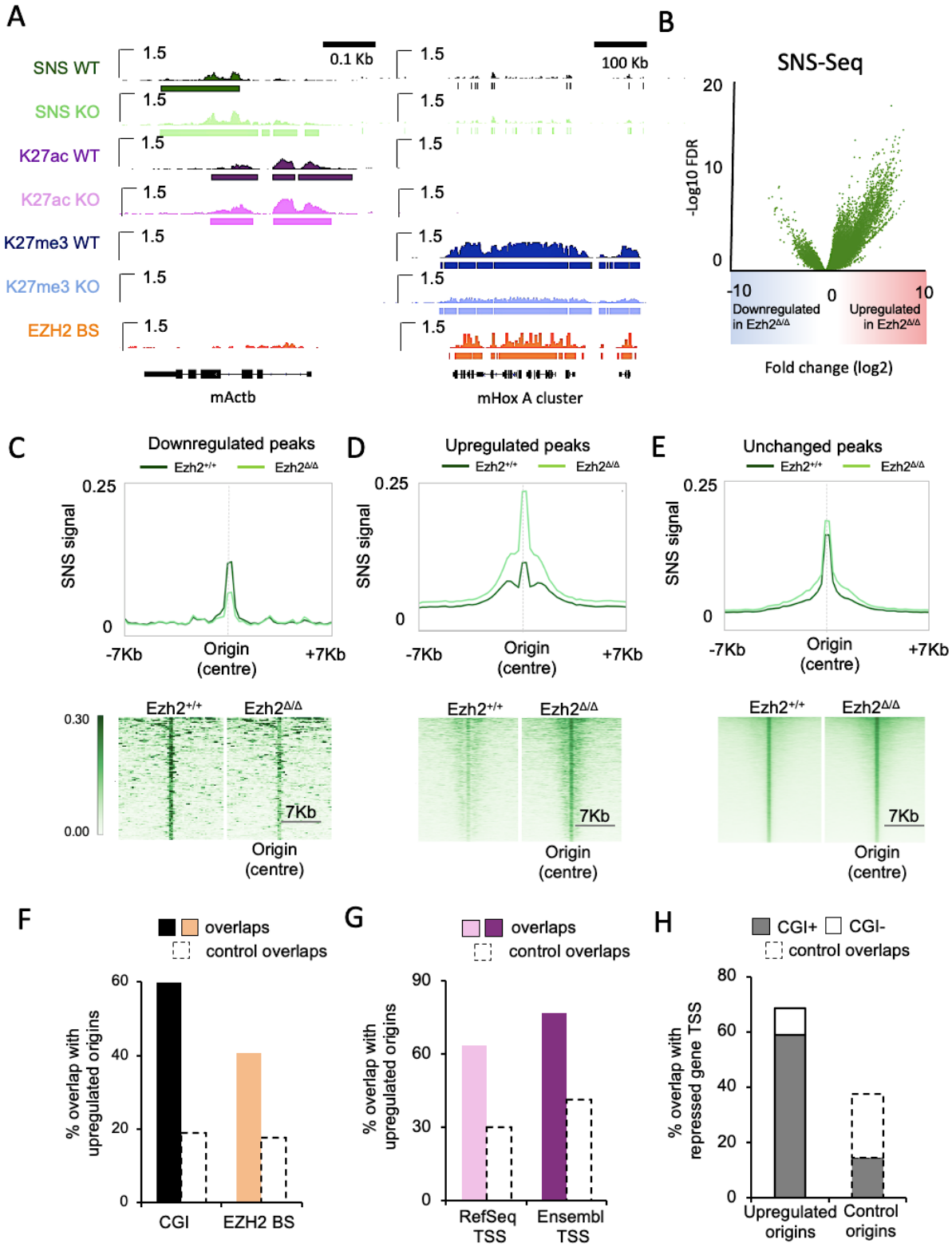


Figure 7. Loss of EZH2 catalytic activity results in increased DNA replication initiation. (A) UCSC genome browser snapshots of SNS-Seq, ChIP-Seq (H3K27me3 and H3K27ac) and EZH2 b nd ng s tes (EZH2 BS) n the nd cated mouse chromosoma coord nates, mActb gene and mHox A c uster. n = 3 ndependent ce cu ture rep cates were performed and merged. **(B)** Vo cano p ot: for each or g n, the corrected p-va ues (fa se d scovery rates, - og$_{10}$(FDR)) and the og$_2$ fo d change (Log2Fo d) of Ezh2$^{+/+}$ and Ezh2$^{\Delta/\Delta}$ samp es were p otted. Or g ns were

c ass f ed nto three d fferent c asses, accord ng to the r act v ty upon EZH2 dep et on: downregu ated, unchanged and upregu ated. **(C-E)** (Top) Read p e up p ots d sp ay ng the s gna prof es from Ezh2$^{+/+}$ and Ezh2$^{\Delta/\Delta}$ ce s centered on downregu ated (113), upregu ated (4,839) and unchanged (23,776) or g ns, respect ve y (±7 Kb). (Bottom) Heatmaps of the nd cated exper ments d sp ay ng s gna prof es from Ezh2$^{+/+}$ and Ezh2$^{\Delta/\Delta}$ ce s centered on downregu ated (113), upregu ated (4,839) and unchanged (23,776) or g ns, respect ve y (±7 Kb). **(F)** Bar p ots represent ng the percentage of over ap between CGI and EZH2 b nd ng s tes (EZH2 BS, 10,114) and upregu ated or g ns (4,839). **(G)** Bar p ots represent ng the percentage of over ap between RefSeq and Ensemb promoters (TSS) and upregu ated or g ns (4,839). **(H)** Bar p ots represent ng the percentage of over ap between upregu ated or g ns (4,839) and contro or g ns (a set of or g ns d sp ay ng unchanged act v ty) and repressed genes w th or w thout CGI promoters (TSS). Upregu ated and downregu ated peaks refer to Ezh2$^{\Delta/\Delta}$ peaks n compar son to Ezh2$^{+/+}$ peaks. The dashed bars show contro over aps w th random zed (shuff ed) reg ons.

The initial findings suggest that loss of PRC2 activity does not lead to a loss of DNA replication activity, on the contrary, it leads to an augmentation of DNA replication activity, specifically at EZH2-bound gene promoters. I next asked whether the chromatin landscape shaped by EZH2 and/or the transcriptional program controlled by EZH2 is involved in this increase in replication origin activity.

4.1.3. Replication initiation and transcriptional de-repression in Ezh2$^{\Delta/\Delta}$ cells

Previous reports suggested a correlation between active transcription and DNA replication initiation [113, 195]. I thus investigated whether upregulation of replication initiation at repressed genes is linked to loss of transcriptional repression. To do this, RNA-Sequencing (RNA-Seq) analysis on mESC cells upon EZH2 inactivation in n = 3 independent replicates was performed. This was followed by differential gene expression (DGE) analysis between the transcriptome of Ezh2$^{+/+}$ and Ezh2$^{\Delta/\Delta}$ using DESeq2. As previously reported by C.F. Pereira et al. [159], Ezh2$^{\Delta/\Delta}$ cells display mild transcriptional dysregulation compared to Ezh2$^{+/+}$ cells. Hence, the volcano plot in Figure 8A shows a similar amount of upregulated and downregulated genes upon EZH2 inactivation (Supplementary Table 1, 524 up and 526 down genes, p-value < 0.05).

Most of the significantly upregulated genes upon EZH2 loss-of-function contain an EZH2 binding site at their promoters (54%, 255 genes, Figure 8B) and some contain DNA replication origins that are also upregulated (23%, 97 genes, Figure 8C), suggesting that loss of transcriptional repression at a minority of promoters could be linked to upregulation of some DNA replication initiation activity in Ezh2$^{\Delta/\Delta}$ cells. However, they represent a minority of EZH2-bound genes.

Despite a mild loss of repression at some gene promoters, the overwhelming majority of EZH2-bound genes remain transcriptionally silent (> 94%, Figure 8D), suggesting that loss-of-function of EZH2 does not result in transcriptional activation for the great majority of PRC-repressed genes. Moreover, of the ~5000 regions that display increased origin activity in $Ezh2^{\Delta/\Delta}$ cells, only a small fraction (< 3%) fell on transcriptionally upregulated genes at TSS or gene body (Figure 8E, F, respectively), suggesting that *de-novo* mRNA transcription is not responsible for the prominent upregulation of replication initiation at repressed gene promoters in $Ezh2^{\Delta/\Delta}$ cells. Consistently, I found that EZH2-bound promoters remain transcriptionally repressed following EZH2 inactivation (Figure 8G) but are associated with a remarkable upregulation of TSS-associated replication origin activity (SNS-Seq, Figure 8H).

I thus conclude that transcriptional de-repression does not underlie an increase in origin activity at repressed promoters upon EZH2 loss.

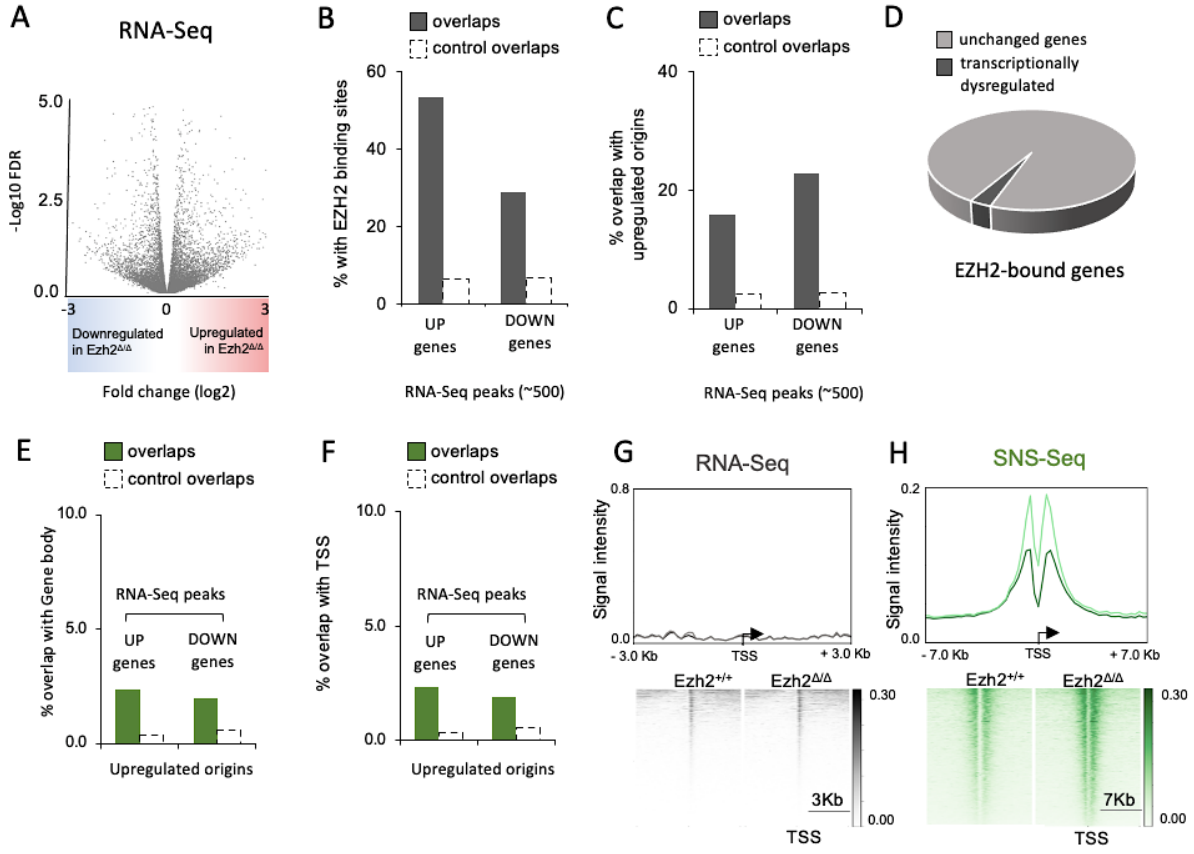


Figure 8. Replication initiation and transcriptional de-repression in $Ezh2^{\Delta/\Delta}$ cells. (A) Vo cano p ot representat on of d fferent a gene express on (RNA-Seq s gna) between $Ezh2^{\Delta/\Delta}$ ce s and $Ezh2^{+/+}$ ce s. The corrected p-va ues (fa se d scovery rates, - og_{10}(FDR)) and the og_2 fo d change (Log2Fo d) of $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ samp es were p otted. **(B, C)** Bar p ots represent ng the percentage of over ap between upregu ated and downregu ated RNA-Seq genes (524 and 526, respect ve y) w th EZH2 b nd ng s tes (10,114) and upregu ated

or g ns (4,839), respect ve y. **(D)** P e chart represent ng the percentage of EZH2-bound genes wh ch are transcr pt ona y unchanged or dysregu ated upon oss of EZH2 nact vat on. (**E, F)** Bar p ots represent ng the percentage of over ap between upregu ated or g ns (4,839) w th the gene body and the TSS of upregu ated and downregu ated RNA-Seq genes (524 and 526, respect ve y), respect ve y. **(G, H)** (Top) Read p e up p ots d sp ay ng the s gna prof es of RNA-Seq and SNS-Seq (act ve or g ns), respect ve y, from $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ ce s, centered on EZH2-bound promoters (± 3 and ±7 Kb, respect ve y). (Bottom) Heatmaps of the nd cated exper ments d sp ay ng s gna prof es assoc ated w th RNA-Seq and SNS-Seq genom c reg ons from $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ ce s, centered on EZH2-bound promoters (± 3 and ±7 Kb, respect ve y). Upregu ated and downregu ated peaks refer to $Ezh2^{\Delta/\Delta}$ peaks n compar son to $Ezh2^{+/+}$ peaks. The dashed bars show contro over aps w th random zed (shuff ed) reg ons.

4.1.4. Increased origin activity in $Ezh2^{\Delta/\Delta}$ cells correlates with EZH2 binding sites

Given that transcriptional de-repression from PRC-repressed genes does not underlie increased DNA replication initiation observed in $Ezh2^{\Delta/\Delta}$ cells, I next examined the chromatin landscape regulated by EZH2 in the context of DNA replication initiation, specifically at EZH2-bound sites. We, therefore, examined the chromatin landscape in $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ cells by chromatin immunoprecipitation (ChIP-Seq) of the relevant histone marks H3K27me3 and H3K27ac. Moreover, I downloaded and re-analyzed publicly available data from two separated EZH2 ChIP-Seq datasets to generate a stringent set of consistent EZH2 binding sites in mouse ESC (Table 2) [165, 166].

Figure 9 displays the signal of the different ChIP-Seq datasets on EZH2-bound regions in read pile up plots. Figure 9A shows how the EZH2 binding sites signal falls in the center of the peaks. EZH2-bound sites experience a prominent upregulation of DNA replication initiation in $Ezh2^{\Delta/\Delta}$ cells (SNS-Seq, Figure 9B). As expected, I noted a similarly prominent loss of deposition for the repressive mark H3K27me3 in $Ezh2^{\Delta/\Delta}$ cells in comparison to $Ezh2^{+/+}$ cells (Figure 9C). The reduction in the repressive H3K27me3 mark was accompanied by small increase in H3K27ac (Figure 9D), a mark that is not directly deposited by PRC, but by CBP and p300 and other regulatory complexes in the cell [196]. Importantly, D. Pasini et al. and others [196-198] have previously reported that global loss of H3K27me3 mark results in unspecific “spreading” of H3K27ac mark, which is consistent with my observations. Control read pile up plots drawn on randomized (shuffled) regions show no variations in EZH2 binding sites, SNS-Seq, H3K27me3 and H3K27ac signal intensity or shape (Supplementary Figure 3I-L, respectively).

A large proportion (>40%) of statistically significant upregulated origins fall on direct binding sites for EZH2 (Figure 7F). Figure 9 shows that direct binding sites of EZH2 experience a loss

of H3K27me3 signal and increased DNA replication initiation, but little to no changes in H3K27ac levels.

Taken together, my results suggest that a loss of H3K27me3 mark, triggered by loss-of-function of EZH2 is associated with DNA replication initiation at EZH2-bound sites in mESC.

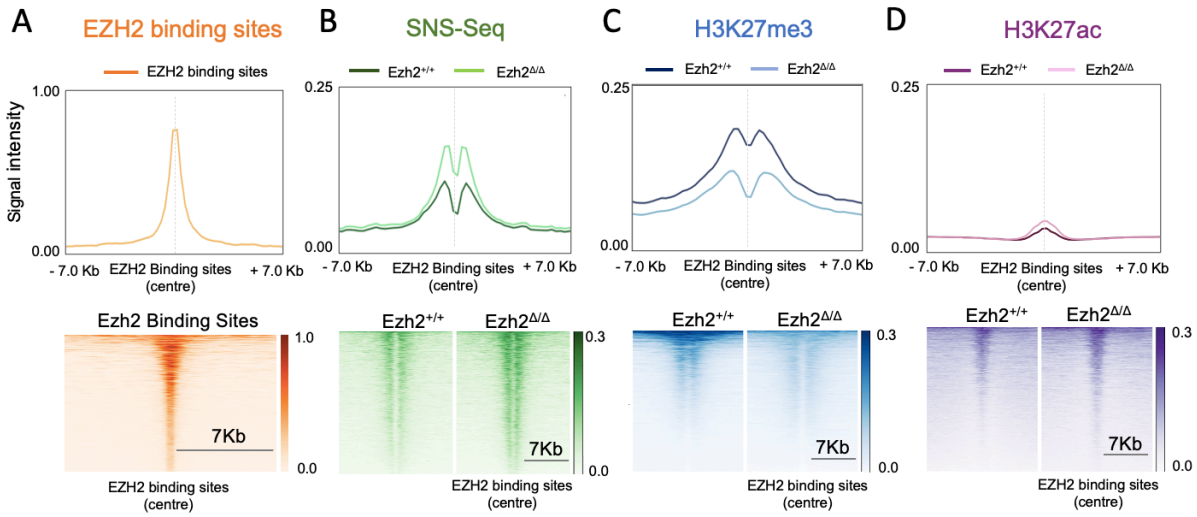


Figure 9. **Increased origin activity in Ezh2$^{\Delta/\Delta}$ cells correlates with EZH2 binding sites. (A-D)** (Top) Read p e up p ots d sp ay ng the s gna prof es of EZH2 b nd ng s tes, SNS-Seq (act ve or g ns), H3K27me3 (repress ve mark) and H3K27ac (act vat ng mark), respect ve y, from Ezh2$^{+/+}$ and Ezh2$^{\Delta/\Delta}$ ce s centered on EZH2 b nd ng s tes peaks (10,114) (±7 Kb). (Bottom) Heatmaps of the nd cated exper ments d sp ay ng s gna prof es from Ezh2$^{+/+}$ and Ezh2$^{\Delta/\Delta}$ ce s, centered on EZH2 b nd ng s tes peaks (10,114) (±7 Kb).

4.1.5. Upregulation in origin activity is linked to changes in chromatin landscape

The upregulation of DNA replication initiation activity appears to be correlated with changes in the chromatin landscape rather than changes in the transcriptional landscape. To study this in detail, I examined the chromatin landscape changes in Ezh2$^{+/+}$ and Ezh2$^{\Delta/\Delta}$ cells concerning DNA replication initiation quantitatively. For that, H3K27me3 and H3K27ac ChIP-Seq was performed in Ezh2$^{+/+}$ and Ezh2$^{\Delta/\Delta}$ cells.

Ezh2$^{+/+}$ cells contain approximately ~14,000 broad domains marked by H3K27me3 (Supplementary Table 1). Of these, 41% are also bound by the PRC2 component EZH2 (Figure 10A), which implies a direct correlation with the loss of methylase activity of EZH2. Furthermore, gene ontology analysis on the H3K27me3 marked regions using GREAT

(Genomic Regions Enrichment of Annotations Tool), reveals developmental biological processes related to PRC2 functions, such as Central Nervous System (CNS) development, confirming a strong correlation between PRC2 activity and the H3K27me3 mark under Prof. M. Méchali's Lab experimental conditions (Figure 10B).

As expected, and previously reported [159, 197, 198], the lack of catalytic activity of EZH2 led to a significant decrease in H3K27me3 in numerous genomic regions (Figure 7A). As seen in Figure 7A, the deposition of the H3K27me3 is significantly reduced at numerous genomic regions as well as PRC2 target loci, where there is a substantial decrease in H3K27me3 in the known PRC2 target, the mHox A cluster (developmental repressed cluster). The loss of H3K27me3 mark we observe is likely to be an underestimation, as quantification by ChIP-Seq fails when one sample has a drastic loss of the precipitated mark (see discussion). I further performed differential peak intensity analysis with Diffbind, which quantitatively supported what was visually observed in the genome browser. As observed in Figure 10C, the volcano plot showed significant downregulation of H3K27me3 mark deposition upon differential peak intensity between $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ cells. Only 39 regions containing H3K27me3 mark are upregulated in $Ezh2^{\Delta/\Delta}$ cells (Supplementary Table 1, FDR < 0.05) and these regions have weak H3K27me3 signal in both between $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ cells (Figure 10D). In contrast, 419 H3K27me3 regions were downregulated in $Ezh2^{\Delta/\Delta}$ cells (Supplementary Table 1, FDR < 0.05). Downregulated peaks showed a strong signal in $Ezh2^{+/+}$ cells and were statistically significantly downregulated in $Ezh2^{\Delta/\Delta}$ cells (Figure 10D). This suggested that the stronger H3K27me3 sites are particularly prone to the loss of EZH2 function.

Importantly, of the regions containing statistically significantly downregulated peaks for the repressive H3K27me3 mark in $Ezh2^{\Delta/\Delta}$ cells, almost all overlapped with EZH2 binding sites (96% Figure 10E) and many also contained upregulated DNA replication initiation sites (32%, Figure 10F). Furthermore, regions that lose H3K27me3 mark display gain in DNA replication initiation signal (Figure 10G), showing a correlation between changes in repressive chromatin landscape and DNA replication initiation. In reverse, when I examine the H3K27me3 signal at the upregulated origins, I find that they also display a loss of H3K27me3 signal (Figure 10H).

Taken together, my results suggest that loss of EZH2 catalytic activity triggers upregulation of DNA replication initiation activity, which is tightly associated with a prominent loss of the repressive mark H3K27me3.

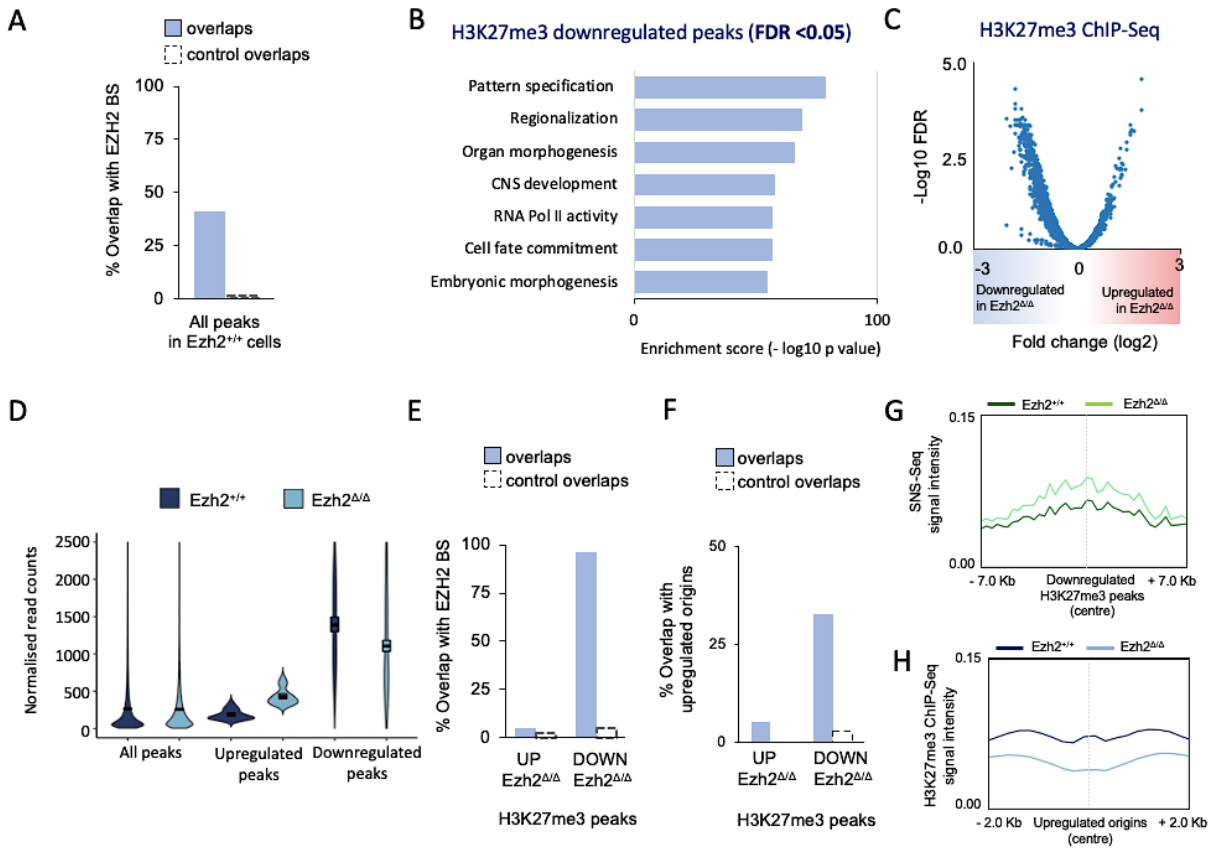


Figure 10. **EZH2 inactivation triggers a loss of H3K27me3 mark and an upregulation of DNA replication initiation activity. (A)** Bar p ots represent ng the percentage of over ap between a H3K27me3 peaks from Ezh2$^{+/+}$ ce s (15,374) and EZH2 b nd ng s tes (10,114). **(B)** Gene onto ogy ana ys s on H3K27me3 downregu ated reg ons. Resu ts of gene onto ogy ana ys s carr ed out us ng the Genom c Reg ons Enr chment of Annotat ons Too (GREAT) on ne b o nformat cs resources. CNS: Centra Nervous System. **(C)** Vo cano p ot representat on of d fferent a H3K27me3 ChIP-Seq s gna between Ezh2$^{Δ/Δ}$ ce s and Ezh2$^{+/+}$ ce s. The corrected p-va ues (fa se d scovery rates, - og$_{10}$(FDR)) and the og$_2$ fo d change (Log2Fo d) of Ezh2$^{+/+}$ and EZH2$^{Δ/Δ}$ samp es were p otted. **(D)** V o n p ot for og2- norma zed read counts of a , upregu ated and downregu ated H3K27me3 peaks (from eft to r ght). The centra part of each "v o n" corresponds to a standard boxp ot, a marker for the med an; symmetr c curves at the s des dep ct the probab ty dens ty. **(E, F)** Bar p ots represent ng the percentage of over ap between upregu ated and downregu ated H3K27me3 peaks (39 and 419 respect ve y) and EZH2 b nd ng s tes (10,114) and upregu ated or g ns (4,839), respect ve y. **(G)** Read p e up p ots d sp ay ng the s gna prof es of SNS-Seq (act ve or g ns) from Ezh2$^{+/+}$ and Ezh2$^{Δ/Δ}$ ce s centered on the 419 downregu ated H3K27me3 peaks (±7 Kb). **(H)** Read p e up p ots d sp ay ng the s gna prof es of H3K27me3 from Ezh2$^{+/+}$ and Ezh2$^{Δ/Δ}$ ce s centered on the 4,839 upregu ated or g ns (±7 Kb). Upregu ated and downregu ated peaks refer to Ezh2$^{Δ/Δ}$ peaks n compar son to Ezh2$^{+/+}$ peaks. The dashed bars show contro over aps w th random zed (shuff ed) reg ons.

I then wished to investigate the effect of not having active PRC2 on the deposition of the activating histone mark H3K27ac and how this related to changes in the DNA replication initiation activity.

EZH2 inactivation produced a genome-wide increase in the signal of the activating histone mark H3K27ac, as visually observed in purple in the genome browser sample in Figure 11A. Differential peak activity analysis of H3K27ac peaks provided a significant number of upregulated peaks upon EZH2 inactivation (~7000 peaks, Supplementary Table 1, Figure 11B). However, quantitatively most of the upregulated peaks had low signal (regions of weak signal) (Figure 11C). Moreover, regions displaying upregulation in H3K27acetylation levels were not restricted to EZH2-bound regions, as less than 10% of all H3K27ac peaks and statistically significantly upregulated H3K27ac regions contained an EZH2 binding site (Figure 11D). Neither did the regions that lost in H3K27me3 mark (downregulated H3K27me3 regions) upon EZH2 inactivation correlate with regions that gained in H3K27ac mark, as the overlap between down and upregulated H3K27me3 and H3K27ac, respectively, was less than 10% (Figure 11E, F). Consistently, the upregulation of the H3K27ac landscape was not associated with biological processes involved in developmental or PRC-mediated functions (Figure 11G), suggesting an indirect effect of EZH2 inactivation on the unspecific increase of H3K27ac mark in chromatin genome-wide. These results were consistent with previous findings by D. Pasini et al. and others [196-198], which described that loss of H3K27me3 mark results in unspecific accumulation of H3K27ac mark genome-wide. Finally, only ~6% of upregulated H3K27ac regions contain significantly upregulated DNA replication origins, which emphasizes the lack of correlation between the gain in DNA replication initiation activity and the unspecific and weak gain in H3K27ac mark.

I conclude that the inactivation of EZH2 triggers a prominent reduction of the repressive mark H3K27me3 at EZH2-bound regions and a spreading of unspecific and new upregulated weak H3K27ac peaks, as reported previously [196-198].

Taken together, the changes in the repressive H3K27me3 (and not H3K27ac) mark landscape triggered by the inactivation of EZH2 created an environment that enables increased DNA replication initiation activity.

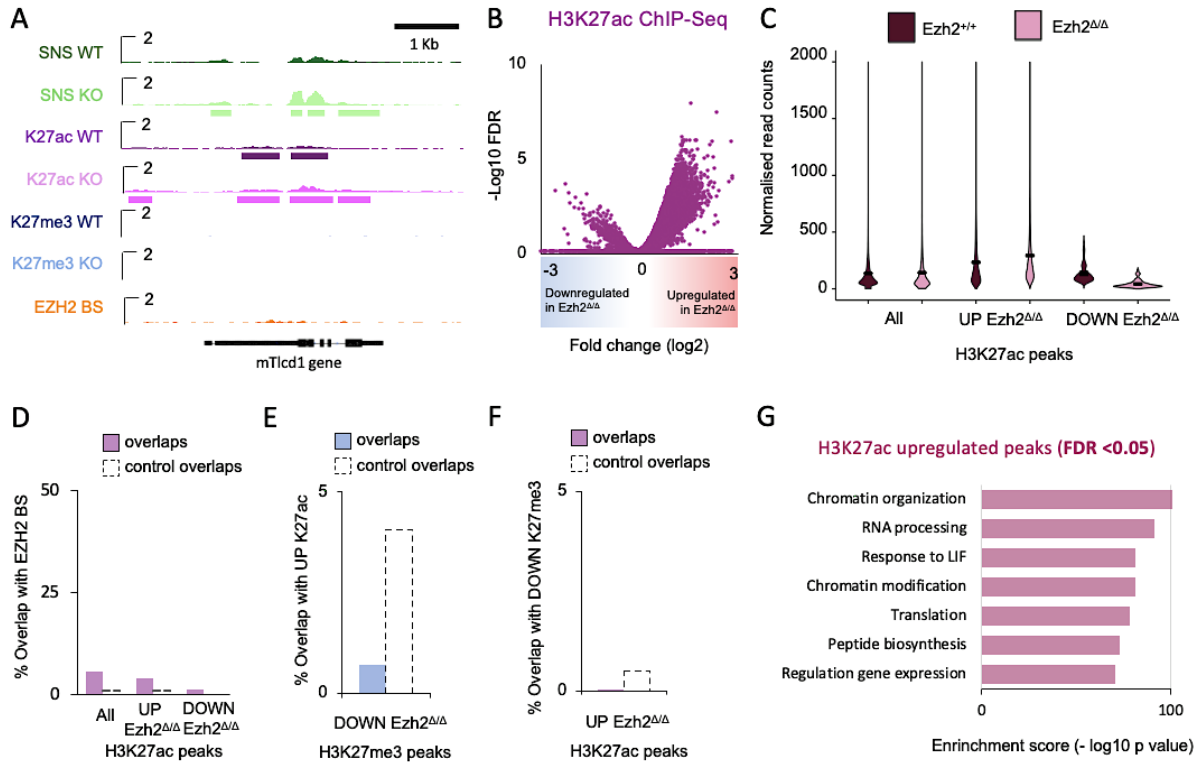


Figure 11. **EZH2 inactivation triggers a spreading of new upregulated H3K27ac peaks. (A)** UCSC genome browser snapshots of SNS-Seq, ChIP-Seq exper ments (H3K27me3 and H3K27ac) and EZH2 b nd ng s tes (EZH2 BS) n a representat ve reg on (mT cd1 gene). Chromosoma coord nates of the sequenc ng tracks are nd cated. n = 3 ndependent ce cu ture rep cates were performed and showed s m ar resu ts. **(B)** Vo cano p ot representat on of d fferent a H3K27ac ChIP-Seq s gna between Ezh2$^{\Delta/\Delta}$ ce s and Ezh2$^{+/+}$ ce s. The corrected p-va ues (fa se d scovery rates, - og$_{10}$(FDR)) and the og$_2$ fo d change (Log2Fo d) of Ezh2$^{+/+}$ and EZH2$^{\Delta/\Delta}$ samp es were p otted. **(C)** V o n p ot for og2- norma zed read counts of a , upregu ated and downregu ated H3K27ac peaks (from eft to r ght). The centra part of each "v o n" corresponds to a standard boxp ot, a marker for the med an; symmetr c curves at the s des dep ct the probab ty dens ty. (**D)** Bar p ots represent ng the percentage of over ap between a H3K27ac peaks from Ezh2$^{+/+}$ ce s (24,553), upregu ated and downregu ated H3K27ac peaks (7,157 and 113 respect ve y) and EZH2 b nd ng s tes (10,114). **(E)** Bar p ots represent ng the percentage of over ap between downregu ated H3K27me3 peaks (419) and upregu ated H3K27ac peaks (7,157). **(F)** Bar p ots represent ng the percentage of over ap between upregu ated H3K27ac peaks (7,157) and downregu ated H3K27me3 peaks (419). **(G)** Gene onto ogy ana ys s on H3K27ac upregu ated reg ons. Resu ts of gene onto ogy ana ys s carr ed out us ng the Genom c Reg ons Enr chment of Annotat ons Too (GREAT) on ne b o nformat cs resources. LIF: Leukem a nh b tory factor. Upregu ated and downregu ated peaks refer to Ezh2$^{\Delta/\Delta}$ peaks n compar son to Ezh2$^{+/+}$ peaks. The dashed bars show contro over aps w th random zed (shuff ed) reg ons.

4.1.6. Upregulation in origin activity is linked to changes in chromatin topological states

The data showed that an increase in replication origin activity at EZH2-binding sites strongly correlated with a corresponding decrease in the repressive H3K27me3 mark. Nonetheless, I

did not find transcriptional de-repression at EZH2-bound promoters. I therefore hypothesized that the transcription-independent upregulation of replication origin activity was linked to the changes in chromatin environment shaped by loss of the repressive H3K27me3 mark. Within a cell, the changes in the histone mark landscape are translated into changes in chromatin topological states. Thus, I expected that the loss in the H3K27me3 mark drives a local increase in open chromatin in EZH2-bound regions. Of note, DNA replication initiation activity is linked to open chromatin regions [14, 16, 19, 25, 43, 46, 48]. Significantly, loss of EZH2 catalytic activity led to an augmentation of DNA replication initiation activity, also in EZH2-bound regions. Therefore, I next investigated the changes in chromatin conformation states in EZH2-bound regions as the potential cause of why there is an increase in DNA replication initiation activity in EZH2-bound regions upon loss of EZH2 function.

To study this hypothesis, I complemented the analysis of the changes in histone epigenetic marks with the analysis of the chromatin topological states. Chromatin topological states (and chromatin accessibility) of the epigenome can be mapped genome-wide using several sequencing techniques, such as ATAC-Seq [75, 94]. While Prof. M. Méchali's Lab did not produce open chromatin maps in the mESC cell lines used for the data displayed in this thesis, I was able to re-analyze publicly available ATAC-Seq data from $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ multipotent cells (LSK murine hematopoietic stem cells, Table 2). Even though there are biological differences between different cell lines, a large number of EZH2 binding sites are shared between similar cell types, especially two pluripotent cell types [199, 200]. ATAC-Seq data from multipotent progenitor cells suggest that EZH2-bound regions in the mESC cell lines utilized in this thesis maintain some open chromatin conformation in $Ezh2^{+/+}$ cells (Figure 12A). Significantly, the signal intensity levels of the ATAC-Seq data on PRC2 repressed regions were similar to the levels experienced by low-activity transcribing genes (Figure 12B, C). I next examined the ATAC-Seq signal upon loss-of-function of EZH2. Z. Gu et al. [167] use a similar strategy to knockout EZH2 function in multipotent progenitor cells, based on genomic recombination by Cre-Lox [168]. When re-analyzed, their data suggests an increase in open chromatin upon loss of Ezh2 activity at EZH2-bound sites (Figure 12D).

Moreover, to validate the changes in open chromatin content (data obtained from murine multipotent hematopoietic stem cells) in the mESC cell lines utilized in this thesis, I also downloaded H3K27me3 and H3K4me3 ChIP-Seq data from the same laboratory, produced with the same cell line alongside ATAC-Seq data (LSK), from which I derived bivalent domains (16,059 peaks, regions containing both the repressive H3K27me3 and activating H3K4me3 marks) (Table 2, Supplementary Table 1) [168]. As described before, promoters repressed by PRC retain bivalency in mESC, a feature that is related to developmental regulation [89-91].

ATAC-Seq data (LSK cells) on bivalent domains from multipotent progenitor cells (LSK cells) suggest an increase in open chromatin in the range of the profile obtained by plotting the same ATAC-Seq data on EZH2 binding sites obtained from mESC (Figure 12D, E). Consistently, the same bivalent domains experienced a decrease of signal in H3K27me3 mark in $Ezh2^{\Delta/\Delta}$ LSK cells, which was in the same signal range of the datasets utilized in this thesis (Figure 9C, Figure 12F). Finally, I only found a slight decrease in the activating H3K4me3 mark in bivalent domains in $Ezh2^{+/+}$ LSK cells, meaning these changes could have not been responsible for the increase in open chromatin in EZH2-inactivated cells. (Figure 12G). Control read pile up plots drawn on randomized (shuffled) regions do not show variations in ATAC-Seq signal intensity or shape (Supplementary Figure 4).

I thus conclude that the changes in the chromatin H3K27me3 mark (but not H3K4me3 mark) landscape upon EZH2 loss were translated into a gain of open chromatin at direct binding sites of EZH2.

EZH2 inactivation produced a nearby chromatin environment where an increase in open chromatin was induced to the levels of low-activity expressed genes. These results suggest that PRC2 indirectly regulates the recruitment of DNA replication initiation activity by repressing nearby chromatin via depositing the repressive H3K27me3 mark in EZH2-bound regions.

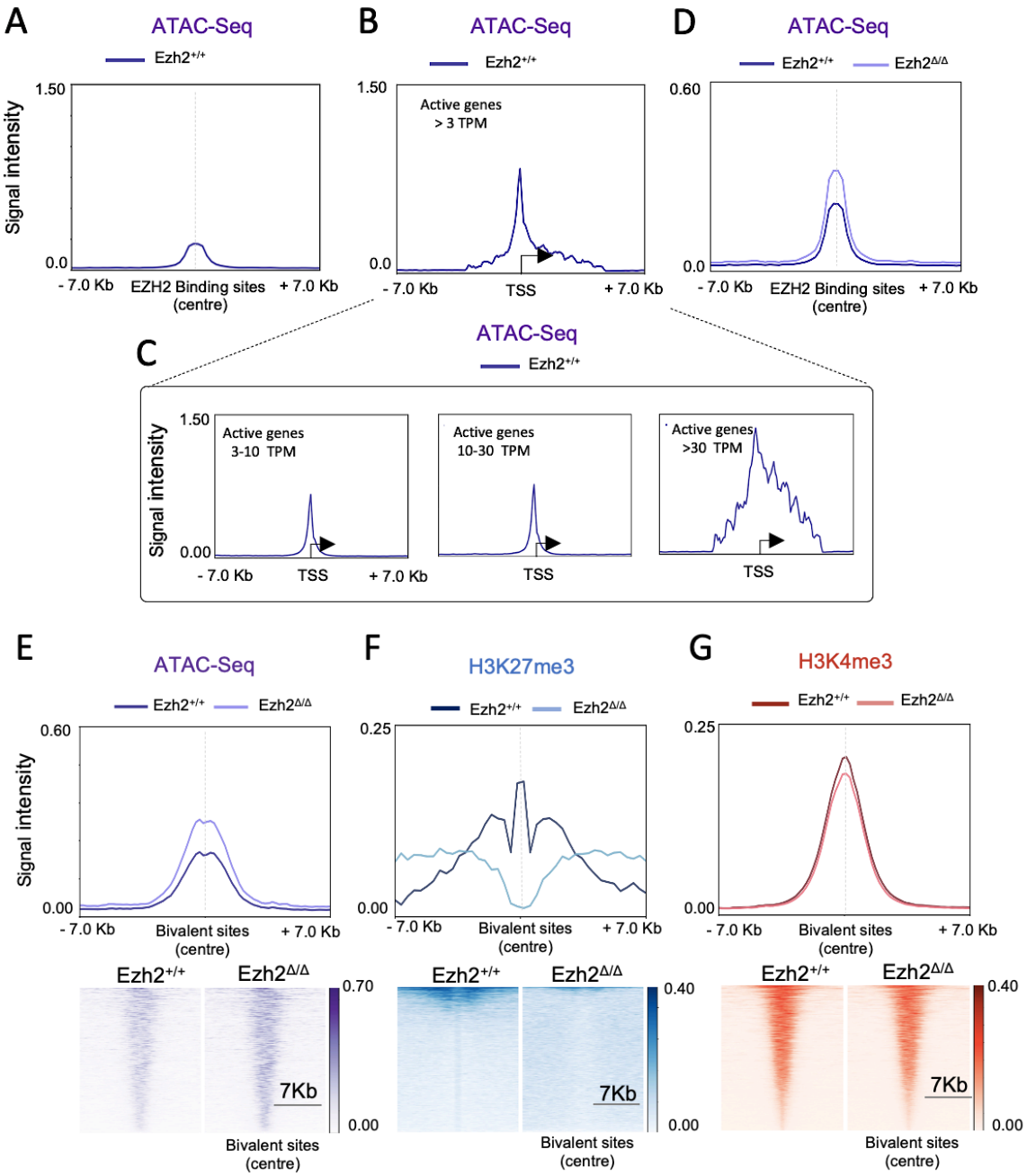


Figure 12. EZH2 inactivation triggers changes in the chromatin topological states. (A) Read p e up p ots d sp ay ng the ATAC-Seq s gna prof e from Ezh2$^{+/+}$ mu t potent ce s centered at 10,114 EZH2 b nd ng s tes peaks (± 7 Kb). **(B)** Read p e up p ots d sp ay ng the ATAC-Seq s gna prof e from Ezh2$^{+/+}$ mu t potent ce s centered at the TSS of transcr pt ona y act ve genes (± 7 Kb). Act ve genes were def ned as hav ng > 3 transcr pts per m on, TPM, n th s ce ne. **(C)** Read p e up p ots d sp ay ng the ATAC-Seq s gna prof e from Ezh2$^{+/+}$ mu t potent ce s centered at the TSS of 3-10 TPM, 10-30 TPM and > 30 TPM transcr pt ona y act ve genes, respect ve y (± 7 Kb). **(D)** Read p e up p ots d sp ay ng the ATAC-Seq s gna prof es from Ezh2$^{+/+}$ and Ezh2$^{Δ/Δ}$ mu t potent ce s centered at 10,114 EZH2 b nd ng s tes peaks (± 7 Kb). **(E-G)** (Top) Read p e up p ots d sp ay ng the ATAC-Seq, H3K27me3 ChIP-Seq and H3K4me3 ChIP-Seq from Ezh2$^{+/+}$ and Ezh2$^{Δ/Δ}$ mu t potent ce s centered on 16,059 B va ent doma n peaks (± 7 Kb). (Bottom) Heatmaps of the nd cated exper ments d sp ay ng s gna prof es from Ezh2$^{+/+}$ and Ezh2$^{Δ/Δ}$ ce s centered on 16,059 B va ent doma n peaks (± 7 Kb). Datasets down oaded from Z. Gu et a ., performed w th the use of LSK mur ne hematopo et c stem ce s [167].

4.2. PART II

While most of my thesis is based on bioinformatic analysis and the impact of the loss of EZH2 function on DNA replication initiation, I have performed wet-lab experiments to study how DNA replication might in turn impact transcription and gene regulation. Unfortunately, this part of my work was disrupted by the ongoing pandemic and remains incomplete.

4.2.1. The deletion of an OGRE/G4 element led to a less active replication origin

DNA replication and transcription co-localize in space and time. Consequently, it is likely that the cell confers some coordination, and probably some cross-effects upon dysregulation, between the two biological processes [26, 115]. I wished to investigate the impact of local gene expression upon the deletion of a single origin and, consequently, Pre-RC complex assembly failure. In this part of the thesis, I began to investigate the potential impact of DNA replication initiation and/or activity on transcription. Even though the work remains unfinished, some of my findings point to a potential regulatory function of DNA replication in transcription.

Here, I used a NIH 3T3 mouse cell line in which a sequence containing a G-quadruplex motif (OGRE/G4 motif), found in the DNA replication origin 1 (Ori1) is mutated ($3T3^{\Delta G4}$, unconditional knockout mouse embryo fibroblast cell line generated by P. Prorok et al. [41] using CRISPR-Cas9 gene editing). Supplementary Figure 2A show the genomic mutation of $3T3^{\Delta G4}$ cells on the mouse Rai1 gene promoter. The matching wild-type line was also received (here referred to as $3T3^{WT}$ cells).

The binding of the origin recognition complex (ORC) to DNA replication origins is a critical event in the initiation of eukaryotic DNA replication [31, 201]. Once bound to DNA, ORC leads the formation of the Pre-RC complex into origins. ORC preferentially binds to the topological element G-quadruplex in Origin G-rich Repeated Elements (what is known as OGRE/G4) [41], an element found in 80-90% of the origins in human and mouse cells at 250-300 bp upstream the initiation site (IS) where the Pre-RC is formed [41, 49]. As reported by P. Prorok et al., the deletion of the G-quadruplex sequence led to a significant decrease in the activity of Ori1 compared to other non-mutated origins (e.g. Ori2) [41]. Interestingly, the researchers reported no impact on nearby gene expression (TSS of Rai1) after the deletion of the OGRE/G4 motif of the endogenous origin Ori1. Methods used to measure gene expression may miss small differences in gene expression. Moreover, Rai1 is a gene regulated by the circadian rhythm [146], therefore, de-synchronization of the circadian rhythm between asynchronous cell

populations (i.e. $3T3^{WT}$ and $3T3^{\Delta G4}$ cells) can significantly impact gene expression. I thus wished to re-study the direct impact of loss of origin activity on transcription of the nearby genes Rai1, Pemt and Srebf-1 using a more sensitive assay and under strictly controlled cell cycle.

4.2.2. Reduced Rai1 gene expression in asynchronous $3T3^{\Delta G4}$ cells

I first wished to determine the impact of deletion of the G-quadruplex motif from Ori1 on the expression of local genes. I quantitatively examined gene expression levels for Rai1, Pemt, and Srebf-1 from asynchronous (AS) $3T3^{WT}$ and $3T3^{\Delta G4}$ cells (Figure 13A-C, from left to right). In the cell lines that were utilized, only the shorter isoform of Rai1 is expressed (i.e. the only active promoter of Rai1 is the one adjacent to the G4 mutation, Supplementary Figure 2A). Interestingly, I find statistically significant differences ($p = 0.005$) in the Rai1 transcriptional activity being lower in $3T3^{\Delta G4}$ by a 35% compared to $3T3^{WT}$ cells (Figure 13A). Interestingly, the activity of the neighboring gene, Srebf-1, (< 10 Kb downstream of Rai1), was also mildly decreased in $3T3^{\Delta G4}$ compared to $3T3^{WT}$ cells (10%, $p = 0.044$) (Figure 13B). Pemt (TSS located ~60 Kb upstream of Rai1 TSS) did not show statistically significant changes ($p > 0.05$) in gene expression activity between $3T3^{WT}$ and $3T3^{\Delta G4}$ cells (Figure 13C).

Hence, the deletion of the G-quadruplex motif, which caused the loss the DNA replication origin activity of Ori1 [41], affected specifically the Rai1 gene promoter, leading to a significant decrease in the transcriptional activity. These results are consistent with two plausible scenarios for how loss-of-function of OGRE/G4 motif may have impacted the expression of Rai1; (i) The deletion of the G-quadruplex of Ori1 resulted in the absence of Pre-RC formation and this defect results in decreased transcription. In this scenario, Pre-RC forms by the direct recruitment of ORC in the OGRE/G4 motif located on the promoter of Rai1, promoting transcription in wild-type cells. In $3T3^{\Delta G4}$ cells, Pre-RC cannot assemble during G1-phase, and this leads to less transcription in AS cells. (ii) Another possibility is that the deletion of the G-quadruplex structure removes both a DNA replication origin but also a functional transcriptional activator sequence. In this scenario, the deletion of the G-quadruplex motif from Ori1 results in decreased transcription due to a deleted cis-regulatory element. The possibility of the G-quadruplex structure acting as a cis-regulatory element has been previously described by others [202].

In order to discern between these two possibilities, I studied the transcriptional activity of Rai1 in cells synchronized in G0-phase of the cell cycle.

4.2.3. The G-quadruplex motif acts as a positive cis-regulatory element

Figure 13D-F shows the Tbp-normalized expression of Rai1, Srebf-1 and Pemt (from right to left) in $3T3^{WT}$ and $3T3^{\Delta G4}$ cells synchronized in G0-phase by serum starvation for 48h and 24h (G0-48h and G0, respectively). Serum starvation causes the cells to exit the cell cycle and, thus, no Pre-RC is formed. Similar to the trends seen in AS cells, Figure 13D also shows that in G0-48h synchronized $3T3^{\Delta G4}$ cells, the expression levels of Rai1 were also statistically significantly decreased in comparison to the expression levels seen in $3T3^{WT}$ cells (40%, $p = 0.013$). As seen in Figure 13E-F, no statistically significant changes in the expression levels of Srebf-1 and Pemt were found on G0-48h or G0 ($p > 0.05$). My observations were sufficient to refuse possibility (i), that Pre-RC formation at the Rai1 promoter upregulates transcription in $3T3^{WT}$ cells.

Consequently, my results support possibility (ii) that the G-quadruplex motif itself functions as a positive cis-regulatory element that promotes Rai1 gene transcription and its absence results in lower transcriptional activity of the Rai1 gene promoter.

However, the fact that the G-quadruplex motif is part of the promoter does not imply that this sequence is not involved in the Pre-RC formation and that Pre-RC formation may or may not impact gene transcription. Thus, I next wished to examine the effects on the transcriptional activity at the time of formation of Pre-RC in synchronously cycling cells.

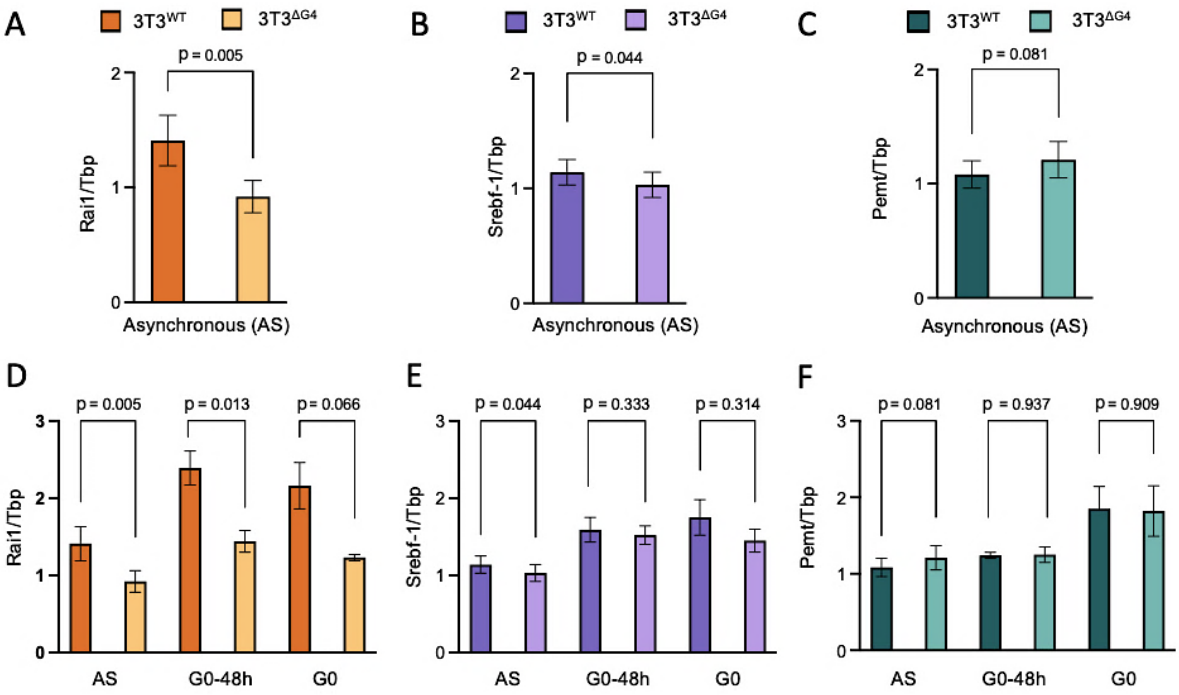


Figure 13. Gene expression levels of Rai1, Pemt and Srebf-1 in AS, G0-48h and G0 synchronized 3T3^{WT} and 3T3$^{\Delta G4}$ cells. (A-C) Bar graphs show ng the re at ve gene express on of Ra 1, Srebf1 and Pemt (r ght to eft) n asynchronous (AS) 3T3^{W} and 3T3$^{\Delta G4}$ ce s measured by RT-Q-PCR. **(D-F)** Bar graphs show ng the re at ve gene express on of Ra 1, Srebf1 and Pemt (r ght to eft) n asynchronous (AS) 3T3^{W} and 3T3$^{\Delta G4}$ ce s and G0-phase synchron zed 3T3^{W} and 3T3$^{\Delta G4}$ ce s by serum starvat on for 48h (G0-48h) and 24h (G0), respect ve y, measured by RT-Q-PCR. Bar he ghts nd cate the nd cate mean gene express on of the genes n 3T3^{W} and 3T3$^{\Delta G4}$ ce s norma zed to TATA-Box B nd ng Prote n gene (Tbp) transcr pt ona act v ty. The p-va ues were ca cu ated us ng a pa red student T-test stat st ca test and nd cate the d fference between the means of the 3T3^{W} and 3T3$^{\Delta G4}$ ce s to 5%. Error bars nd cate the SEM. n = 5.

4.2.4. Pre-RC formation interferes with adjacent gene expression

Next, I determined the impact of Pre-RC formation during G1-phase on nearby transcription. Previous studies that used cell sorting and ORC1 quantification on chromatin suggest that following release from serum starvation, NIH 3T3 cells start to form Pre-RC within 3-8h, followed by entry into S-phase at 10-14h [203, 204]. I examined gene expression in cells entering G1-phase (2h, 4h, 6h, 8h post-release), followed by S-phase (16h and 18h post-release) simultaneously from both 3T3^{WT} and 3T3$^{\Delta G4}$ cells. Flow cytometry analysis (FACS) was carried out to determine the phase of the cell cycle the cells were in at different time points (Figure 14A). As confirmed with FACS analysis, both 3T3^{WT} and 3T3$^{\Delta G4}$ cells synchronize in G0/G1 upon serum starvation and release back into the cell cycle with similar kinetics, where the majority of cells have entered and progressed through S-phase by 16h (Figure 14A).

As shown in Figure 14B in green, the gene expression levels of Pemt in 3T3$^{\Delta G4}$ cells remain similar to the gene expression levels of 3T3^{WT} cells levels throughout the cell cycle (3T3$^{\Delta G4}$ / 3T3^{WT} gene expression ratio close to 1). Unlike Pemt gene expression, there is a difference in the gene expression levels of Rai1 in 3T3$^{\Delta G4}$ cells compared to 3T3^{WT} cells at G0 to 4h with an average of a ~40% less Rai1 mRNA in 3T3$^{\Delta G4}$ cells (Figure 14B in orange). However, the transcriptional activity of Rai1 can be seen to increase at the 4-6h time point and reaches maximum gene expression at the 8h time point (Figure 14B). The expression levels of Rai1 in 3T3$^{\Delta G4}$ cells at the 8h time point almost recovers to wild-type expression levels as there is only a ~20% difference in the gene expression between the two cell genotypes (Figure 14B). Once the cells enter S-phase, at 16 and 18h time points, Rai1 gene expression in 3T3$^{\Delta G4}$ cells returns to the basal levels observed in G0-phase synchronized cells with a ~40% lower Rai1 mRNA in 3T3$^{\Delta G4}$ cells compared to 3T3^{WT} cells (Figure 14B).

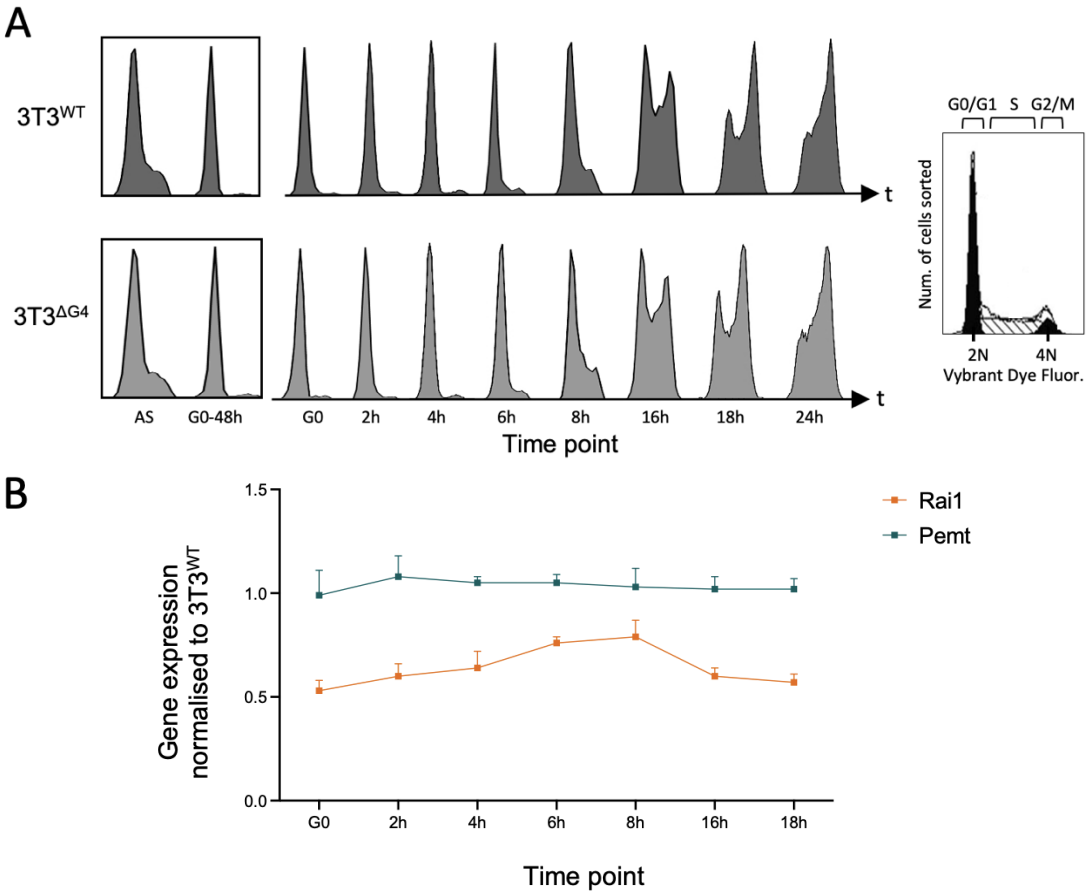


Figure 14. **FACS analysis and normalized gene expression of 3T3$^{\Delta G4}$ cells post-release**. **(A)** F ow cytometry ana ys s (FACS) was used to determ ne the ce DNA content, wh ch corre ates w th the ce cyc e phases. (R ght) Representat ve h stogram of F ow cytometry ana ys s (FACS) of NIH 3T3 ce s show ng the re at ve f uorescence ntens ty of Vybrant Dye vs the number of ce s sorted. 2N and 4N refer to the number of cop es of DNA n the ce . G0/G1-phase ce s are found n the 2N peak and G2/M-phase ce s are found n the 4N peak. Ce s trans t ng from G1 to G2 (DNA synthes s phase, S-phase) are found n between the two peaks. The eft f gures show FACS of (top) 3T3^{W} and (bottom) 3T3$^{\Delta G4}$ ce s from G0-phase synchron zed ce s and re eased ce s a ong the course of the exper ment (2h, 4h, 6h, 8h, 16h and 18h). **(B)** L ne p ot show ng the re at ve gene express on eve s of Ra 1 and Pemt n cyc ng ce s across G0, 2h, 4h, 6h, 8h, 16h and 18h t me po nts n 3T3$^{\Delta G4}$ ce s. Leve s are f rst norma zed to Tbp and then to w d-type eve s. Error bars nd cate the SEM. n = 5.

My experiment suggests that in time points where Pre-RC is not yet formed (G0, 2h, 4h, referred to as “pre”) Rai1 transcriptional activity in 3T3$^{\Delta G4}$ cells is ~40% lower compared to 3T3^{WT} cells (Figure 15A, left bar). Rai1 mRNA levels rise during the time period of Pre-RC formation (6h, 8h, referred to as Pre-RC), and this increase is statistically significantly ($p = 0.005$) (Figure 15A, middle bar). Moreover, once the Pre-RC disassembles upon S-phase entry (16h, 18h, referred to as “post”) relative Rai1 mRNA levels go back to pre – Pre-RC levels (Figure 15A, right bar). These results confirmed that the G-quadruplex element functions as a positive cis-regulatory element. Moreover, the results suggest that formation of

Pre-RC complex at the G-quadruplex structure or in the vicinity of the Rai1 promoter was inhibitory to transcription in $3T3^{WT}$ cells around the time of Pre-RC formation, and only during this time period.

Srebf-1 display slightly lower expression trends in $3T3^{\Delta G4}$ cells compared to $3T3^{WT}$ cells at all three stages, although the changes are non-significant (Figure 15B). This suggests that deleting a G-quadruplex structure from an adjacent promoter may affect the transcription levels of nearby genes [60, 144, 205]. There were also non-significant changes ($P > 0.05$) in the transcriptional activity of Pemt (whose promoter is located at ~ 60 Kb upstream the deletion) and the non-local control gene Actb amongst the three stages of Pre-RC formation, which $3T3^{\Delta G4}$ / $3T3^{WT}$ gene expression ratio was found to be ~1 (Figure 15C, D). Of note, the promoters of Pemt, Srebf-1 and Actb genes do not contain DNA replication origins (data not shown).

My results are consistent with previous findings that some G-quadruplex elements are essential structures which, when located in promoters, could function as positive cis-regulatory elements. Pre-RC complexes bind and form on G-quadruplex. **Taken together, my results suggest that when Pre-RCs are formed upstream the Rai1 promoter, the formation of the complex can specifically inhibit the transcriptional activity of the Rai1 gene but not flanking neighboring genes, and only during a dedicated period during G1-phase.**

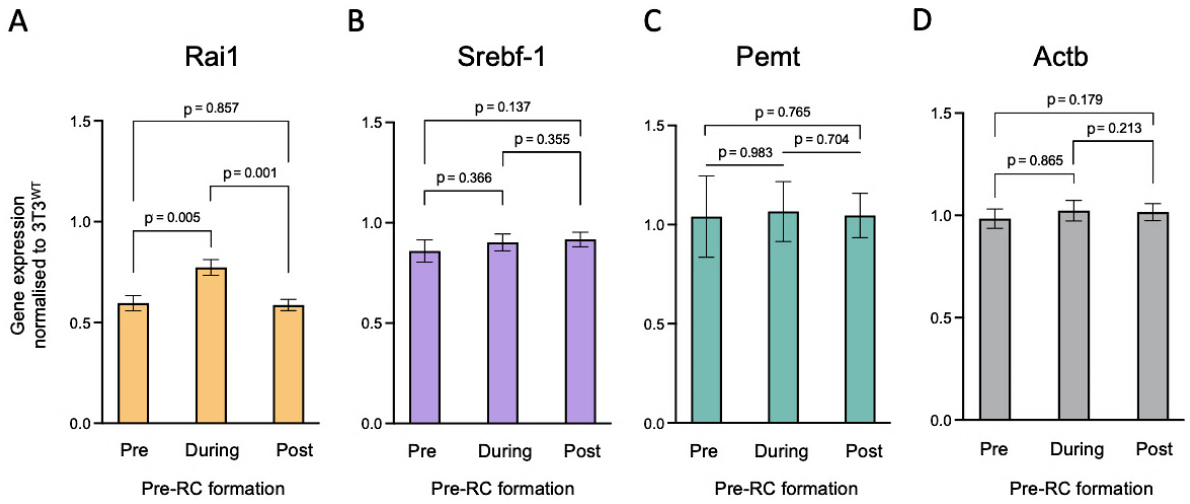


Figure 15. **Normalized gene expression of $3T3^{\Delta G4}$ cells at different stages of the Pre-RC formation. (A-D)** The figure shows bar plots displaying the relative $3T3^{\Delta G4}$ cells gene expression levels of Rai1, Srebf-1, Pemt and Actb, from left to right, before Pre-RC formation (Pre, G0/2/4h), at the time of Pre-RC formation (during, 6/8h) and post Pre-RC formation (post, 16/18h). The p-values were calculated using a paired student T-test statistical test and indicate the difference between the means of the three groups (pre, during and post Pre-RC formation) to 5%. Error bars indicate the SEM. n = 5.

5. DISCUSSION & CONCLUSIONS

5.1. PART I: Re-shaped chromatin environment recruits DNA replication initiation upon loss-of-function of EZH2

The positioning of DNA replication origins throughout the genome is not random. 79% of EZH2-bound silent genes contain DNA replication origins at their promoters (Figure 6C). Based on these preliminary data, I hypothesized that PRC2 may be involved in the active recruitment of DNA replication initiation activity at silent gene promoters. To test this hypothesis, I comprehensively characterized the consequences of EZH2 depletion on the activity of DNA replication origins in mESC through a genome-wide evaluation of $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ cells via high-throughput sequencing.

If EZH2, or its catalytic activity, was directly involved in recruiting DNA replication initiation to PRC-bound sites, the absence of EZH2 catalytic activity would fail to recruit initiation activity to these sites. However, my results show that abrogation of EZH2 (or its catalytic activity) led to a prominent increase in DNA replication activity (~5000 peaks, Figure 7B), specially at CGI-containing, EH2-bound silent promoters (Figure 7F and 9B). There could be two plausible explanations for such an increase in DNA replication initiation activity in EZH2-inactivated cells; de-repression of the transcriptional program regulated by EZH2 and/or chromatin landscape re-shaped by EZH2 loss.

Transcription has been positively linked to DNA replication initiation activity [113, 195]. However, I encountered that the great majority of EZH2-bound promoters remained transcriptionally silent (Figure 8). I thus conclude that active mRNA transcription is not responsible for the increased DNA replication initiation at EZH2-bound promoters in $Ezh2^{\Delta/\Delta}$ cells. My results also indirectly suggest that DNA replication initiation activity can be de-coupled from active transcription. Given the presented findings, more research is warranted on the role of (if any) active transcription on recruitment of DNA replication initiation.

Since I did not find transcription to underlie an increase in DNA replication initiation, I next investigated the changes in the histone epigenetic mark landscape resulting from depletion of EZH2 catalytic activity in the context of DNA replication initiation, specifically at EZH2-bound sites. My data is consistent with previous observations that loss of PRC2/EZH2 catalytic activity leads to a prominent loss of the repressive mark H3K27me3 (Figure 9C and 10E), and a further unspecific genome-wide spreading of the activating mark H3K27ac (Figure 9D and 11) [196-198]. While DNA replication activity has been liked to H3K27ac, I do not find a significant amount of DNA replication initiation associated with the spreading of this mark (only

~ 6% of upregulated H3K27ac peaks contain upregulated origins), suggesting perhaps that the small amount of spreading of this activating mark on its own cannot recruit DNA replication activity, or that this “spreading” signal is too weak to cause the necessary conformational changes to recruit origins.

On the other hand, the regions containing statistically significantly downregulated peaks for the H3K27me3 mark in $Ezh2^{\Delta/\Delta}$ cells co-localized with EZH2 binding sites (96% of overlap, Figure 10E). Furthermore, regions that lose H3K27me3 mark display gain in the DNA replication initiation signal (Figure 10G), showing a strong correlation between changes in this repressive mark landscape and DNA replication initiation. Importantly, in a cell, the changes in histone mark landscape are translated into changes in chromatin topological states. Therefore, as PRC2 inactivation led to changes in the landscape of repressive and activating marks (H3K27me3 and H3K27ac, respectively), I hypothesized that PRC2 inactivation triggers changes in the content of open chromatin, which in turn may recruit DNA replication initiation**.** To test this hypothesis, I investigated the changes in the topological states in EZH2-bound regions and nearby chromatin upon loss of EZH2 function in a publicly available dataset.

Despite the lack of transcriptional de-repression, PRC-repressed promoters maintain a small amount of open chromatin to the levels comparable to low-activity expressed genes, and this open chromatin at PRC-bound sites is increased upon EZH2 inactivation (Figure 12). Open chromatin has previously been associated with DNA replication initiation activity and may underlie high DNA replication initiation activity at these PRC-repressed promoters [21]. Hence, I suggest that PRC-repressed promoters remain transcriptionally silent but acquire more open chromatin upon loss of EZH2 function due to a loss in the repressive H3K27me3 mark. Consequently, an increase in open chromatin confers a chromatin environment where DNA replication initiation activity is recruited, specifically in PRC-bound regions. Interestingly, my analysis suggests that the H3K4me3 mark, an activating mark usually deposited at promoters does not significantly increase, suggesting that changes to H3K27me3 are more likely behind the relaxation/opening of chromatin.

Given similar cell cycle dynamics between the $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ cells (Supplementary Figure 1G), **I conclude that the inactivation of the PRC2 catalytic activity protein EZH2 does not directly recruit DNA replication initiation activity, and the loss of EZH2-meditated H3K27me3 mark creates changes in the chromatin environment near PRC-repressed genes favouring DNA replication initiation. Importantly, this model cannot rule out the possibility that the physical presence of PRC, which can still assemble at its binding sites may help recruit DNA replication initiation activity.**

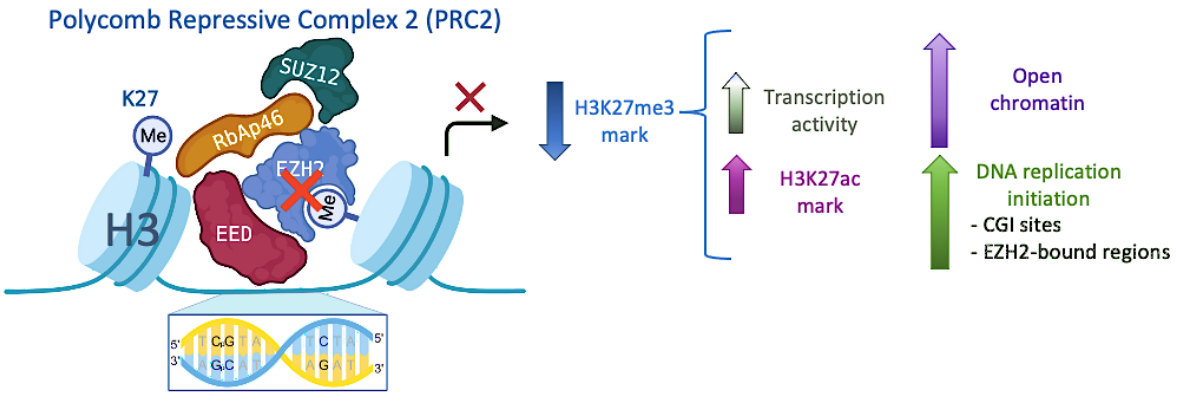


Figure 16. Model suggested in this thesis by which PRC2 recruits DNA replication initiation activity. My findings suggest that the loss of the catalytic activity of PRC and/or the related epigenetic mark (H3K27me3) creates an environment that favours DNA replication initiation, specifically at CGI sites, EZH2-bound regions. Figure made with Biorender.com

5.1.1. Limitations and future work

My work provides evidence on the mechanism by which the Polycomb complex proteins 2 (PRC2) regulates the recruitment of DNA replication initiation origins. However, my results present some limitations owing to the designed experimental plan and availability of experimental methods.

Concerning ChIP-Seq, I encountered limitations regarding the available sequencing techniques. Importantly, I only retrieved ~400 peaks downregulated H3K27me3 peaks under the experimental conditions of Prof. M. Méchali Lab. However, I believe the real loss in H3K27me3 is much more drastic at many more sites, which I have not been able to detect [72]. Although a semi-quantitative instrumental technique, ChIP-Seq data is challenging when mapping global changes of histone post-translational modifications and fail to detect the real epigenetic landscape due to a lack of reliable normalizing reference control, as reported by B. Egan et al. with an EZH2 inhibitor system [206]. Moreover, the normalization parameters used by Diffbind differential peak analysis do not account for large changes in peak intensity (i.e. almost all normalization methods used by these programs assume the majority of peaks remain unchanged). To overcome this challenge, some have applied alternative normalization approaches such as Spike-in ChIP-Seq (SpikChIP), which is based on the addition of other specie's chromatin to minimize the overcorrection of genomic regions in ChIP-Seq [206, 207]. I believe the use of SpikChIP, or another convenient normalization approach, would have provided a more accurate ChIP-Seq comparison in the mESC cell lines utilized in this thesis.

I did not perform ChIP-Seq to map the sites where EZH2 binds in the genome. I downloaded publicly available data also performed in mESC, which I consider valid and similar to the experimental conditions of Prof. M. Méchali Lab (Table 2). However, it would have been more accurate to perform EZH2 ChIP-Seq in the mESC cell lines utilized in this thesis under the experimental conditions of Prof. M. Méchali Lab. Similarly, Prof. M. Méchali Lab did also not perform ATAC-Seq and H3K4me3 ChIP-Seq in the mESC cell lines utilized in this thesis, which would have generated more accurate and comprehensive comparisons with the SNS-Seq data displayed in this thesis.

Future work would include the study of the role of other catalytic or functional Polycomb complex proteins from both PRC1 and PRC2, (i.e. EED, SUZ12, RING1A/B) in the recruitment of DNA replication initiation origins. This would elucidate whether the physical presence of the PRC complex at PRC-bound sites plays a role in DNA replication initiation; the model system used in this thesis is based on lack of H3K27me3 mark deposition and cannot discount the possibility that the physical presence of PRC2 may actively play a role in the recruitment of origin activity. Of note, Prof. M. Méchali Lab obtained the same sequencing data presented in this work for an induced knockout mESC cell line for EED [208], another core component of PRC2, which I then analyzed following the same pipeline as used in EZH2 sequencing data. However, due to experimental and/or technical errors, the data was not reliable to use. The experimental replicates should be reproduced to obtain comprehensive data.

5.2. PART II: Pre-RC formation hampers transcription from Rai1 promoter

Due to the co-localization of DNA replication and transcription, and the widespread positioning of origins at transcription promoters rather than non-promoter regions, it is likely that some coordination between the two biological events exists [26, 115]. While many studies investigated the impact of transcription on DNA replication initiation features (activity, choice, and timing), few ever investigated the potential impact of DNA replication initiation positioning on gene regulation. Thus, questions remain unanswered regarding the potential role of DNA replication (initiation) as a gene-regulatory element. For example, are origins mostly located at promoters for a reason? Can DNA replication origins confer an extra layer of transcriptional regulation?

To begin to answer these questions, I used a model system where the NIH 3T3 mouse cell line (mEF) in which a sequence containing a G-quadruplex motif (OGRE/G4 motif) is knockout ($3T3^{\Delta G4}$, Supplementary Figure 2, generated by P. Prorok et al. [41]). As reported by the

researchers, the deletion of such OGRE/G4 motif leads to a decrease in the activity of a closely-located endogenous DNA replication origin (Ori1) [41].

Initial quantitative evaluation of the differential expression of nearby genes in asynchronous (AS) cells showed that the deletion of the G-quadruplex motif affected specifically the expression of Rai1, leading to a statistically significant reduction (35%, $p = 0.005$) in the transcriptional activity in $3T3^{\Delta G4}$ AS cells in comparison to $3T3^{WT}$ AS cells (Figure 13A). Moreover, the activity of the closest gene, Srebf-1, (< 10 Kb downstream) was also mildly decreased in $3T3^{\Delta G4}$ cells (10%, $p = 0.044$, Figure 13B), whilst the upstream gene Pemt remained unaffected (Figure 13C).

These observations are consistent with two plausible scenarios for how loss-of-function of the OGRE/G4 motif may have impacted the expression of Rai1; (i) The first possibility is that the Pre-RC presence on the promoter may promote the transcription of Rai1 during G1-phase. Thus, in $3T3^{\Delta G4}$ cells, the deletion of the G-quadruplex motif impedes the formation of Pre-RC. Without Pre-RC formation, Rai1 gene expression in asynchronous $3T3^{\Delta G4}$ cells is decreased. (ii) A second possibility is that the G-quadruplex sequence functions as a cis-regulatory element that promotes transcription of Rai1. Thus, in asynchronous $3T3^{\Delta G4}$ cells, I observe less Rai1 transcription in the absence of this cis-regulatory element. To distinguish between the two plausible scenarios, I studied the transcription levels of Rai1 in G0-synchronized $3T3^{WT}$ and $3T3^{\Delta G4}$ cells, where no Pre-RC formation takes place. Therefore, if (i) is correct, the transcriptional activity of Rai1 in G0-phase $3T3^{\Delta G4}$ and $3T3^{WT}$ cells should be equivalent. If (ii) is correct, Rai1 gene expression in G0-phase $3T3^{\Delta G4}$ cells relative to $3T3^{WT}$ cells should be lower, akin to the observations in asynchronous cells since the deletion of the G-quadruplex/cis-regulatory element would affect both asynchronous and G0-phase cells.

I concluded that the G-quadruplex motif itself works as a positive cis-regulatory element, which promotes Rai1 transcription and its absence results in lower transcriptional activity of the Rai1 promoter (possibility ii) (Figure 17A).

Aside from its role as a cis-regulatory element, my experiments suggest that the OGRE/G4 motif can impact adjacent transcription at the time of Pre-RC formation (Figure 14B and 15A). The results of my experiments in synchronized cells are consistent with an established role for the OGRE/G4 motif in recruiting Pre-RC components and the Pre-RC complex conferring a mild, transient inhibitory effect on the transcriptional output of a local gene TSS, Rai1 (Figure 14B and 15A). This inhibition is gene-specific and *in cis*, as adjacent gene promoters remain unaffected (Figures 14B and 15B, C). More importantly, the inhibition is restricted to the time of Pre-RC formation and is not observable before Pre-RC formation or after Pre-RC

dissociation (Figure 14B and 15A), suggesting a direct impact of the physical presence of Pre-RC on gene transcription (Figure 17B).

Conflicts, in particular, collisions between the transcription and replication machinery that can impact gene expression have previously been described [116, 117]. These conflicts result in the activation of checkpoint and DNA repair pathways which may in turn impact local gene expression [115, 117, 209]. My observations come from G1-phase synchronized cells (Figure 14A), where no active DNA synthesis or firing of origins occurs. The OGRE/G4 motif and the DNA replication initiation sites are located approximately 250-300 bp upstream of the TSS [41, 49], whereas the actively transcribing RNA pol II would be found at or downstream of the TSS. Consequently, my data does not support a scenario in which transcriptional machinery and the Pre-RC "collision" cause DNA repair pathways to activate which in turn impact nearby gene expression. Instead, my data is consistent with Pre-RC formation at the promoter inhibiting transcription due to steric hinderance between the transcription initiation machinery and the Pre-RC, which are both large, multi-protein complexes.

Many TSSs contain G4 motifs and open chromatin, potentially serving as binding platforms for the sequence-unspecific ORC complex [16]. Transcription initiation is not continuous but likely happens in "bursts" of activity [210, 211]. The binding of ORC to TSS may transiently inhibit or delay the binding of transcription initiation machinery, delaying the next burst of transcriptional activity, resulting in decreased transcriptional output. Alternatively, co-binding of ORC and transcription initiation machinery may cause steric hinderance and hamper adequate transcription initiation, handicapping the bursts of transcriptional activity during Pre-RC formation.

In conclusion, I have demonstrated that a single G-quadruplex (G4) motif found at a DNA replication origin (Ori1) can act as a cis-regulatory element. While the G4 motif was found to function as a transcriptional cis-regulatory element, I also found that the formation of the Pre-RC complex on this G4 element (probably directed by the presence of the G4 motif) had an inhibitory impact on transcription as removal of Ori1 led to increased transcription of Rai1 at the time of Pre-RC formation (and only at the time of Pre-RC formation).

As far as I am concerned, this is the first evidence in mammalian cells that suggests that the Pre-RC formation at promoters have an inhibitory effect on promoter transcription outside of transcription-replication machinery collisions. I hypothesize this biological event as a mechanism of avoidance of steric impedances between replication and transcription

complexes. These results require further investigation but suggest that DNA replication initiation can impact proximal gene transcription. As such, DNA replication origins may act as gene regulatory elements and play part in the pathophysiology of human diseases.

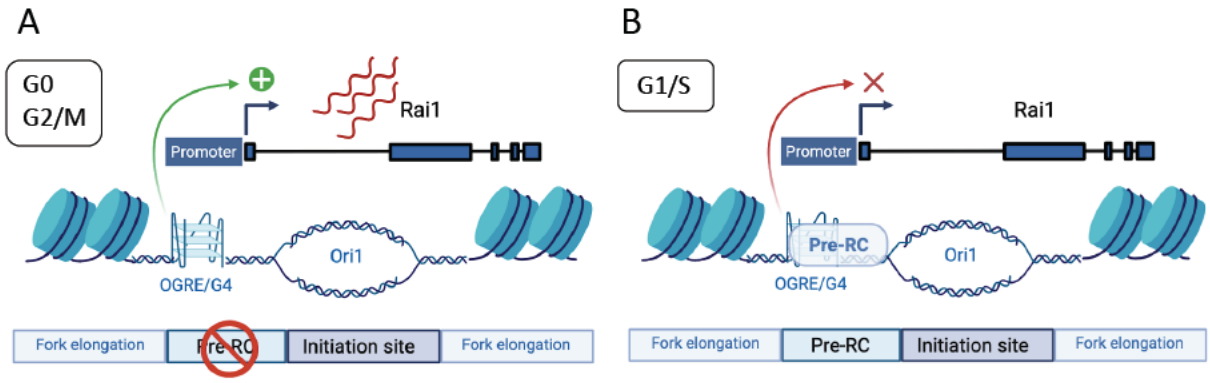


Figure 17. Model suggested in this thesis by which Pre-RC formation inhibits nearby transcription. (A) OGRE/G4 structures can act as c s-regu atory e ements, promot ng transcr pt on of adjacent genes. At the t me when Pre-RC does not form nto the endogenous or g n Or 1 (G0, G2/M -phases), Ra 1 transcr pt on s upregu ated. **(B)** When a Pre-RC comp ex b nd and forms n the OGRE/G4 mot f from Or 1, such format on can confer nh b tory effects on the transcr pt ona output of the adjacent gene Ra 1. F gure made w th B orender.com

5.2.1. Limitations and future work

Future work is necessary to interpret my results correctly and describe the mechanisms by which origin positioning affects gene transcription on a genome-wide scale.

While I have relied on previously published studies to determine the timing of Pre-RC formation [203, 204], this study needs to determine the accurate timing of Pre-RC formation under my experimental conditions. This could be done through the isolation of chromatin followed by western blotting of Pre-RC components, such as ORC1 or CDT1, from $3T3^{WT}$ and $3T3^{\Delta G4}$ cells at the specified time points post-release. Moreover, a more sensitive and quantitative evaluation of the percentages of S-phase cells in each time point by BrdU immunostaining assay or BrdU-sorting is warranted.

I decided to look at the gene expression of Rai1, Srebf-1 and Pemt specifically at 2h, 4h, 6h, 8h, 16h and 18h. The choice of this set of time points allowed us to view a trend in gene expression through the cell cycle and Pre-RC formation, which I grouped the time points into prior, during and post Pre-RC formation. However, there is a gap in my data between 8h and 16h post-release due to the time constraints of this project. During that period origins “fire”,

which is probably when direct collisions between the replicative polymerase and the transcriptional machinery occur. To study the impact on adjacent and near gene expression when origins “fire”, future work should include the addition of more time points in the experimental plan at two-hour intervals between the 8h-16h gap (10h, 12h, 14h post-release). The addition of such time points would enable us to determine the increasing trend by which Rai1 gene expression increases in $3T3^{\Delta G4}$ cells close to $3T3^{WT}$ cell expression, and even if the increase is higher than what was detected with my experimental plan. The addition of more experimental replicates (n = 8-10) would also be desirable in order to account for non-significant variations, as detection of small variations in gene expression is difficult.

Finally, I have studied the impact of the deletion of a single origin on a single gene promoter (and two flanking promoters). Future work could also include the study of other origins that have similar sequence elements to Ori1 (i.e. G-quadruplex, other enhancers and presence of promoters) to validate what was found on Ori1 and the Rai1 promoter as a mechanism that also affects other promoters.

Moreover, to study how the Pre-RC formation impacts transcription genome-wide (on a global scale) and provide information on this, a similar approach to the experimental plan presented in Part I of this thesis should be performed with a Pre-RC component (i.e. ORC mapping by ChIP-seq, origin activity mapping by SNS-Seq and genome-wide transcriptome evaluation by RNA-Seq). Prof. Gustavo Leone Lab generated a Cre-induced knockout embryo fibroblast mouse cell line for ORC1 to study the ORC1 function in genome endoreduplication and organ development [212]. Dr. Akerman Lab received $ORC1^{fl/fl}$ cells as a gift from Prof. Gustavo Leone to investigate the affectation of the transcriptome after de-positioning of origin replication initiation. This project is still ongoing and the experimental samples from $ORC1^{fl/fl}$ cells are due to be sequenced in the University of Birmingham Genomics and Sequencing facility.

6. PUBLICATION OUTCOMES

Manuscript under preparation for EMBO journal:

Title: The Ezh2 Polycomb group protein regulates DNA replication origin firing at Polycomb sites

Authors: Paulina Prorok[1 2#*], Faezeh Forouzanfar[1#], Nerea Murugarren[3 4#], Romain Charton[1], Isabelle Peiffer[1], Ildem Akerman[3 4*] and Marcel Méchali[1*]

Author Information

1. Institute of Human Genetics CNRS-University de Montpellier, France.
2. Present address: Department of Biology, Technical University of Darmstadt, 64287 Darmstadt, Germany.
3. Institute of Metabolism and Systems Research (IMSR), University of Birmingham, UK.
4. Centre for Endocrinology, Diabetes and Metabolism, Birmingham Health Partners, Birmingham, United Kingdom.

[#] Equal author contribution

* Co-corresponding authors

Bibliography

1. Berti, M. and A. Vindigni, *Replication stress: getting back on track.* Nature structural & molecular biology, 2016. **23**(2): p. 103-109.
2. Abbas, T., M.A. Keaton, and A. Dutta, *Genomic instability in cancer.* Cold Spring Harbor perspectives in biology, 2013. **5**(3): p. a012914.
3. Flach, J., et al., *Replication stress is a potent driver of functional decline in ageing haematopoietic stem cells.* Nature, 2014. **512**(7513): p. 198-202.
4. Kurat, C.F., et al., *Chromatin controls DNA replication origin selection, lagging-strand synthesis, and replication fork rates.* Molecular cell, 2017. **65**(1): p. 117-130.
5. Siddiqui, K., K.F. On, and J.F. Diffley, *Regulating DNA replication in eukarya.* Cold Spring Harbor perspectives in biology, 2013. **5**(9): p. a012930.
6. Tanaka, S. and H. Araki, *Helicase activation and establishment of replication forks at chromosomal origins of replication.* Cold Spring Harbor perspectives in biology, 2013. **5**(12): p. a010371.
7. Yeeles, J.T., et al., *Regulated eukaryotic DNA replication origin firing with purified proteins.* Nature, 2015. **519**(7544): p. 431-435.
8. Cayrou, C., et al., *Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features.* Genome research, 2011. **21**(9): p. 1438-1449.
9. Abid Ali, F., et al., *Cryo-EM structure of a licensed DNA replication origin.* Nature communications, 2017. **8**(1): p. 1-10.
10. Miller, T.C. and A. Costa, *The architecture and function of the chromatin replication machinery.* Current opinion in structural biology, 2017. **47**: p. 9-16.
11. Wasserman, M.R., et al., *Replication fork activation is enabled by a single-stranded DNA gate in CMG helicase.* Cell, 2019. **178**(3): p. 600-611. e16.
12. Wang, M.D., et al., *Physics meets biology: The joining of two forces to further our understanding of cellular function.* Molecular Cell, 2021. **81**(15): p. 3033-3037.
13. Sánchez, H., et al., *DNA replication origins retain mobile licensing proteins.* Nature communications, 2021. **12**(1): p. 1-12.
14. Leonard, A.C. and M. Méchali, *DNA replication origins.* Cold Spring Harbor perspectives in biology, 2013. **5**(10): p. a010116.
15. Sequeira-Mendes, J., et al., *Differences in firing efficiency, chromatin, and transcription underlie the developmental plasticity of the Arabidopsis DNA replication origins.* Genome research, 2019. **29**(5): p. 784-797.
16. Akerman, I., et al., *A predictable conserved DNA base composition signature defines human core DNA replication origins.* Nature communications, 2020. **11**(1): p. 1-15.
17. Besnard, E., et al., *Unraveling cell type–specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs.* Nature structural & molecular biology, 2012. **19**(8): p. 837-844.
18. Petryk, N., et al., *Replication landscape of the human genome.* Nature communications, 2016. **7**(1): p. 1-13.
19. Langley, A.R., et al., *Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq).* Nucleic acids research, 2016. **44**(21): p. 10230-10247.
20. Bell, S.P. and K. Labib, *Chromosome duplication in Saccharomyces cerevisiae.* Genetics, 2016. **203**(3): p. 1027-1067.
21. MacAlpine, H.K., et al., *Drosophila ORC localizes to open chromatin and marks sites of cohesin complex loading.* Genome research, 2010. **20**(2): p. 201-211.
22. Kuzminov, A., *Collapse and repair of replication forks in Escherichia coli.* Molecular microbiology, 1995. **16**(3): p. 373-384.
23. Ganier, O., et al., *Metazoan DNA replication origins.* Current opinion in cell biology, 2019. **58**: p. 134-141.

24. de la Paz Sanchez, M., et al., *Regulating DNA replication in plants.* Cold Spring Harbor Perspectives in Biology, 2012. **4**(12): p. a010140.
25. Méchali, M., *Eukaryotic DNA replication origins: many choices for appropriate answers.* Nature reviews Molecular cell biology, 2010. **11**(10): p. 728-738.
26. Fragkos, M., et al., *DNA replication origin activation in space and time.* Nature reviews Molecular cell biology, 2015. **16**(6): p. 360-374.
27. Gilbert, D.M., *Making sense of eukaryotic DNA replication origins.* Science, 2001. **294**(5540): p. 96-100.
28. Huberman, J.A. and A.D. Riggs, *On the mechanism of DNA replication in mammalian chromosomes.* Journal of molecular biology, 1968. **32**(2): p. 327-341.
29. Remus, D. and J.F. Diffley, *Eukaryotic DNA replication control: lock and load, then fire.* Curr Opin Cell Biol, 2009. **21**(6): p. 771-7.
30. Dutta, A. and S.P. Bell, *Initiation of DNA replication in eukaryotic cells.* Annual review of cell and developmental biology, 1997. **13**(1): p. 293-332.
31. Bell, S.P. and B. Stillman, *ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex.* Nature, 1992. **357**(6374): p. 128-134.
32. Bochman, M.L. and A. Schwacha, *The Mcm2-7 complex has in vitro helicase activity.* Molecular cell, 2008. **31**(2): p. 287-293.
33. Li, H. and M.E. O'Donnell, *The eukaryotic CMG helicase at the replication fork: emerging architecture reveals an unexpected mechanism.* Bioessays, 2018. **40**(3): p. 1700208.
34. Douglas, M.E., et al., *The mechanism of eukaryotic CMG helicase activation.* Nature, 2018. **555**(7695): p. 265-268.
35. Ilves, I., et al., *Activation of the MCM2-7 helicase by association with Cdc45 and GINS proteins.* Molecular cell, 2010. **37**(2): p. 247-258.
36. Deegan, T.D. and J.F. Diffley, *MCM: one ring to rule them all.* Current opinion in structural biology, 2016. **37**: p. 145-151.
37. Moyer, S.E., P.W. Lewis, and M.R. Botchan, *Isolation of the Cdc45/Mcm2–7/GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase.* Proceedings of the National Academy of Sciences, 2006. **103**(27): p. 10236-10241.
38. Zou, Y., et al., *Functions of human replication protein A (RPA): from DNA replication to DNA damage and stress responses.* Journal of cellular physiology, 2006. **208**(2): p. 267-273.
39. Lee, S.-H., et al., *Synthesis of DNA by DNA polymerase epsilon in vitro.* Journal of Biological Chemistry, 1991. **266**(33): p. 22707-22717.
40. Alberts, B., *DNA replication and recombination.* Nature, 2003. **421**(6921): p. 431-435.
41. Prorok, P., et al., *Involvement of G-quadruplex regions in mammalian replication origin activity.* Nature communications, 2019. **10**(1): p. 1-16.
42. Gilbert, D.M., *Replication origin plasticity, Taylor-made: inhibition vs recruitment of origins under conditions of replication stress.* Chromosoma, 2007. **116**(4): p. 341-347.
43. Kit Leng Lui, S., et al., *Monitoring genome-wide replication fork directionality by Okazaki fragment sequencing in mammalian cells.* Nature Protocols, 2021. **16**(2): p. 1193-1218.
44. Macheret, M. and T.D. Halazonetis, *Monitoring early S-phase origin firing and replication fork movement by sequencing nascent DNA from synchronized cells.* Nature Protocols, 2019. **14**(1): p. 51-67.
45. Tubbs, A., et al., *Dual roles of poly (dA: dT) tracts in replication initiation and fork collapse.* Cell, 2018. **174**(5): p. 1127-1142. e19.
46. Dellino, G.I., et al., *Genome-wide mapping of human DNA-replication origins: levels of transcription at ORC1 sites regulate origin selection and replication timing.* Genome research, 2013. **23**(1): p. 1-11.
47. Miotto, B., Z. Ji, and K. Struhl, *Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers.* Proceedings of the National Academy of Sciences, 2016. **113**(33): p. E4810-E4819.

48. 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, 2008. **105**(41): p. 15837-15842.
49. Cayrou, C., et al., *The chromatin environment shapes DNA replication origin organization and defines origin classes.* Genome research, 2015. **25**(12): p. 1873-1885.
50. Cayrou, C., et al., *Genome-scale identification of active DNA replication origins.* Methods, 2012. **57**(2): p. 158-164.
51. Hyrien, O., *Peaks cloaked in the mist: the landscape of mammalian replication origins.* Journal of Cell Biology, 2015. **208**(2): p. 147-160.
52. Foss, M., et al., *Origin recognition complex (ORC) in transcriptional silencing and DNA replication in S. cerevisiae.* Science, 1993. **262**(5141): p. 1838-1844.
53. Marahrens, Y. and B. Stillman, *A yeast chromosomal origin of DNA replication defined by multiple functional elements.* Science, 1992. **255**(5046): p. 817-823.
54. Rhind, N. and D.M. Gilbert, *DNA replication timing.* Cold Spring Harb Perspect Biol, 2013. **5**(8): p. a010132.
55. 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, 2013. **23**(11): p. 1774-1788.
56. Deaton, A.M. and A. Bird, *CpG islands and the regulation of transcription.* Genes & development, 2011. **25**(10): p. 1010-1022.
57. Delgado, S., et al., *Initiation of DNA replication at CpG islands in mammalian chromosomes.* The EMBO journal, 1998. **17**(8): p. 2426-2435.
58. Comoglio, F., et al., *High-resolution profiling of Drosophila replication start sites reveals a DNA shape and chromatin signature of metazoan origins.* Cell Rep, 2015. **11**(5): p. 821-34.
59. Xu, J., et al., *Genome-wide identification and characterization of replication origins by deep sequencing.* Genome Biology, 2012. **13**(4): p. R27.
60. Prorok, P., et al., *Involvement of G-quadruplex regions in mammalian replication origin activity.* Nature Communications, 2019. **10**(1): p. 3274.
61. Cayrou, C., et al., *The chromatin environment shapes DNA replication origin organization and defines origin classes.* Genome Res, 2015. **25**(12): p. 1873-85.
62. Cayrou, C., et al., *New insights into replication origin characteristics in metazoans.* Cell Cycle, 2012. **11**(4): p. 658-67.
63. Picard, F., et al., *The spatiotemporal program of DNA replication is associated with specific combinations of chromatin marks in human cells.* PLoS Genet, 2014. **10**(5): p. e1004282.
64. Hoshina, S., et al., *Human origin recognition complex binds preferentially to G-quadruplex-preferable RNA and single-stranded DNA.* J Biol Chem, 2013. **288**(42): p. 30161-30171.
65. Capra, J.A., et al., *G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in Saccharomyces cerevisiae.* PLoS computational biology, 2010. **6**(7): p. e1000861-e1000861.
66. Rawal, P., et al., *Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation.* Genome research, 2006. **16**(5): p. 644-655.
67. Verma, A., et al., *Genome-wide computational and expression analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species.* J Med Chem, 2008. **51**(18): p. 5641-9.
68. Foti, R., et al., *Nuclear Architecture Organized by Rif1 Underpins the Replication-Timing Program.* Molecular Cell, 2016. **61**(2): p. 260-273.
69. Kanoh, Y., et al., *Rif1 binds to G quadruplexes and suppresses replication over long distances.* Nature Structural & Molecular Biology, 2015. **22**(11): p. 889-897.
70. Gnan, S., et al., *Nuclear organisation and replication timing are coupled through RIF1–PP1 interaction.* Nature Communications, 2021. **12**(1): p. 2910.
71. Waddington, C.H., *Organisers and genes.* Organisers and genes., 1940.

72. You, Y., et al., *Temporal dynamics of gene expression and histone marks at the Arabidopsis shoot meristem during flowering.* Nature Communications, 2017. **8**(1): p. 15120.
73. Tsompana, M. and M.J. Buck, *Chromatin accessibility: a window into the genome.* Epigenetics & Chromatin, 2014. **7**(1): p. 33.
74. Klein, D.C. and S.J. Hainer, *Genomic methods in profiling DNA accessibility and factor localization.* Chromosome Res, 2020. **28**(1): p. 69-85.
75. Buenrostro, J.D., et al., *ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide.* Current protocols in molecular biology, 2015. **109**: p. 21.29.1-21.29.9.
76. Miller, J.L. and P.A. Grant, *The role of DNA methylation and histone modifications in transcriptional regulation in humans.* Sub-cellular biochemistry, 2013. **61**: p. 289-317.
77. Allis, C.D. and T. Jenuwein, *The molecular hallmarks of epigenetic control.* Nat Rev Genet, 2016. **17**(8): p. 487-500.
78. Aggarwal, B.D. and B.R. Calvi, *Chromatin regulates origin activity in Drosophila follicle cells.* Nature, 2004. **430**(6997): p. 372-376.
79. Eaton, M.L., et al., *Conserved nucleosome positioning defines replication origins.* Genes & development, 2010. **24**(8): p. 748-753.
80. Santos-Rosa, H., et al., *Methylation of histone H3 K4 mediates association of the Isw1p ATPase with chromatin.* Mol Cell, 2003. **12**(5): p. 1325-32.
81. Rando, O.J., *Global patterns of histone modifications.* Curr Opin Genet Dev, 2007. **17**(2): p. 94-9.
82. Tardat, M., et al., *The histone H4 Lys 20 methyltransferase PR-Set7 regulates replication origins in mammalian cells.* Nat Cell Biol, 2010. **12**(11): p. 1086-93.
83. Ferrari, K.J., et al., *Polycomb-dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity.* Mol Cell, 2014. **53**(1): p. 49-62.
84. Schuettengruber, B., et al., *Genome Regulation by Polycomb and Trithorax: 70 Years and Counting.* Cell, 2017. **171**(1): p. 34-57.
85. Slany, R.K., *When epigenetics kills: MLL fusion proteins in leukemia.* Hematol Oncol, 2005. **23**(1): p. 1-9.
86. Chang, P.Y., et al., *Binding of the MLL PHD3 finger to histone H3K4me3 is required for MLL-dependent gene transcription.* J Mol Biol, 2010. **400**(2): p. 137-44.
87. Kumar, D., et al., *Decoding the function of bivalent chromatin in development and cancer.* Genome Res, 2021. **31**(12): p. 2170-84.
88. Vastenhouw, N.L. and A.F. Schier, *Bivalent histone modifications in early embryogenesis.* Current opinion in cell biology, 2012. **24**(3): p. 374-386.
89. Ku, M., et al., *Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains.* PLoS Genet, 2008. **4**(10): p. e1000242.
90. Bernstein, B.E., et al., *A bivalent chromatin structure marks key developmental genes in embryonic stem cells.* Cell, 2006. **125**(2): p. 315-26.
91. Azuara, V., et al., *Chromatin signatures of pluripotent cell lines.* Nat Cell Biol, 2006. **8**(5): p. 532-8.
92. Mantsoki, A., G. Devailly, and A. Joshi, *CpG island erosion, polycomb occupancy and sequence motif enrichment at bivalent promoters in mammalian embryonic stem cells.* Scientific Reports, 2015. **5**(1): p. 16791.
93. Reverón-Gómez, N., et al., *Accurate Recycling of Parental Histones Reproduces the Histone Modification Landscape during DNA Replication.* Molecular Cell, 2018. **72**(2): p. 239-249.e5.
94. Buenrostro, J.D., et al., *Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.* Nature methods, 2013. **10**(12): p. 1213-1218.
95. Kalantry, S. and T. Magnuson, *The Polycomb group protein EED is dispensable for the initiation of random X-chromosome inactivation.* PLoS genetics, 2006. **2**(5): p. e66-e66.

96. Aoto, T., et al., *Polycomb Group Protein-associated Chromatin Is Reproduced in Post-mitotic G1 Phase and Is Required for S Phase Progression*.* Journal of Biological Chemistry, 2008. **283**(27): p. 18905-18915.
97. Herviou, L., et al., *EZH2 in normal hematopoiesis and hematological malignancies.* Oncotarget, 2015. **7**.
98. Kuzmichev, A., et al., *Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein.* Genes Dev, 2002. **16**(22): p. 2893-905.
99. Cao, R., et al., *Role of histone H3 lysine 27 methylation in Polycomb-group silencing.* Science, 2002. **298**(5595): p. 1039-43.
100. Wang, X., et al., *Targeting of Polycomb Repressive Complex 2 to RNA by Short Repeats of Consecutive Guanines.* Mol Cell, 2017. **65**(6): p. 1056-1067.e5.
101. Margueron, R., et al., *Ezh1 and Ezh2 maintain repressive chromatin through different mechanisms.* Mol Cell, 2008. **32**(4): p. 503-18.
102. Shen, X., et al., *EZH1 mediates methylation on histone H3 lysine 27 and complements EZH2 in maintaining stem cell identity and executing pluripotency.* Mol Cell, 2008. **32**(4): p. 491-502.
103. Czermin, B., et al., *Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites.* Cell, 2002. **111**(2): p. 185-96.
104. van der Vlag, J. and A.P. Otte, *Transcriptional repression mediated by the human polycomb-group protein EED involves histone deacetylation.* Nature Genetics, 1999. **23**(4): p. 474-478.
105. Pasini, D., et al., *Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity.* Embo j, 2004. **23**(20): p. 4061-71.
106. Radulović, V., G. de Haan, and K. Klauke, *Polycomb-group proteins in hematopoietic stem cell regulation and hematopoietic neoplasms.* Leukemia, 2013. **27**(3): p. 523-533.
107. Deevy, O. and A.P. Bracken, *PRC2 functions in development and congenital disorders.* Development (Cambridge, England), 2019. **146**(19): p. dev181354.
108. Wang, H., et al., *Role of histone H2A ubiquitination in Polycomb silencing.* Nature, 2004. **431**(7010): p. 873-8.
109. MacAlpine, D.M., H.K. Rodríguez, and S.P. Bell, *Coordination of replication and transcription along a Drosophila chromosome.* Genes & development, 2004. **18**(24): p. 3094-3105.
110. Kohzaki, H. and Y. Murakami, *Transcription factors and DNA replication origin selection.* Bioessays, 2005. **27**(11): p. 1107-16.
111. Goldman, M.A., et al., *Replication timing of genes and middle repetitive sequences.* Science, 1984. **224**(4650): p. 686-692.
112. Martin, M.M., et al., *Genome-wide depletion of replication initiation events in highly transcribed regions.* Genome research, 2011. **21**(11): p. 1822-1832.
113. Knott, S.R., et al., *Forkhead transcription factors establish origin timing and long-range clustering in S. cerevisiae.* Cell, 2012. **148**(1-2): p. 99-111.
114. Filippova, G.N., et al., *An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes.* Molecular and cellular biology, 1996. **16**(6): p. 2802-2813.
115. Saponaro, M., *Transcription-Replication Coordination.* Life (Basel, Switzerland), 2022. **12**(1): p. 108.
116. Merrikh, H., et al., *Co-directional replication-transcription conflicts lead to replication restart.* Nature, 2011. **470**: p. 554-7.
117. Mirkin, E.V. and S.M. Mirkin, *Mechanisms of transcription-replication collisions in bacteria.* Mol Cell Biol, 2005. **25**(3): p. 888-95.
118. Gagniuc, P. and C. Ionescu-Tirgoviste *Eukaryotic genomes may exhibit up to 10 generic classes of gene promoters*. BMC genomics, 2012. **13**, 512 DOI: 10.1186/1471-2164-13-512.

119. Smale, S.T. and J.T. Kadonaga, *The RNA polymerase II core promoter.* Annu Rev Biochem, 2003. **72**: p. 449-79.
120. Gershenzon, N.I. and I.P. Ioshikhes, *Synergy of human Pol II core promoter elements revealed by statistical sequence analysis.* Bioinformatics, 2005. **21**(8): p. 1295-300.
121. Lagrange, T., et al., *New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB.* Genes & development, 1998. **12**(1): p. 34-44.
122. Yang, C., et al., *Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters.* Gene, 2007. **389**(1): p. 52-65.
123. Saxonov, S., P. Berg, and D.L. Brutlag, *A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters.* Proceedings of the National Academy of Sciences, 2006. **103**(5): p. 1412-1417.
124. Jabbari, K. and G. Bernardi, *Cytosine methylation and CpG, TpG (CpA) and TpA frequencies.* Gene, 2004. **333**: p. 143-9.
125. Bird, A., *DNA methylation patterns and epigenetic memory.* Genes Dev, 2002. **16**(1): p. 6-21.
126. Wilusz, J.E., H. Sunwoo, and D.L. Spector, *Long noncoding RNAs: functional surprises from the RNA world.* Genes & development, 2009. **23**(13): p. 1494-1504.
127. Weingarten-Gabbay, S., et al., *Systematic interrogation of human promoters.* Genome research, 2019. **29**(2): p. 171-183.
128. Blackwood, E.M. and J.T. Kadonaga, *Going the distance: a current view of enhancer action.* Science, 1998. **281**(5373): p. 60-3.
129. Pennacchio, L.A., et al., *Enhancers: five essential questions.* Nature reviews. Genetics, 2013. **14**(4): p. 288-295.
130. Schoenfelder, S. and P. Fraser, *Long-range enhancer–promoter contacts in gene expression control.* Nature Reviews Genetics, 2019. **20**(8): p. 437-455.
131. Martin, M.M., et al., *Genome-wide depletion of replication initiation events in highly transcribed regions.* Genome research, 2011. **21**(11): p. 1822-1832.
132. Hiratani, I., et al., *Global reorganization of replication domains during embryonic stem cell differentiation.* PLoS Biol, 2008. **6**(10): p. e245.
133. Hiratani, I. and D.M. Gilbert, *Replication timing as an epigenetic mark.* Epigenetics, 2009. **4**(2): p. 93-97.
134. Raghuraman, M.K., et al., *Replication dynamics of the yeast genome.* Science, 2001. **294**(5540): p. 115-21.
135. Heichinger, C., et al., *Genome-wide characterization of fission yeast DNA replication origins.* Embo j, 2006. **25**(21): p. 5171-9.
136. Müller, C.A. and C.A. Nieduszynski, *DNA replication timing influences gene expression level.* Journal of Cell Biology, 2017. **216**(7): p. 1907-1914.
137. Di Rienzi, S.C., et al., *Maintaining replication origins in the face of genomic change.* Genome Res, 2012. **22**(10): p. 1940-52.
138. van der Meijden, C.M., et al., *Gene profiling of cell cycle progression through S-phase reveals sequential expression of genes required for DNA replication and nucleosome assembly.* Cancer Res, 2002. **62**(11): p. 3233-43.
139. Hereford, L.M., et al., *Cell-cycle regulation of yeast histone mRNA.* Cell, 1981. **24**(2): p. 367-75.
140. Omberg, L., et al., *Global effects of DNA replication and DNA replication origin activity on eukaryotic gene expression.* Molecular systems biology, 2009. **5**: p. 312.
141. Williamson, L., et al., *UV Irradiation Induces a Non-coding RNA that Functionally Opposes the Protein Encoded by the Same Gene.* Cell, 2017. **168**(5): p. 843-855.e13.
142. Tanaka, S. and J.F. Diffley, *Deregulated G1-cyclin expression induces genomic instability by preventing efficient pre-RC formation.* Genes Dev, 2002. **16**(20): p. 2639-49.
143. Macheret, M. and T.D. Halazonetis, *Intragenic origins due to short G1 phases underlie oncogene-induced DNA replication stress.* Nature, 2018. **555**(7694): p. 112-116.

144. Varshney, D., et al., *The regulation and functions of DNA and RNA G-quadruplexes.* Nature Reviews Molecular Cell Biology, 2020. **21**(8): p. 459-474.
145. Pennuto, M., et al., *Chapter 9 - Induced pluripotent stem cells for modeling Smith–Magenis syndrome*, in *Current Progress in iPSC Disease Modeling*, A. Birbrair, Editor. 2022, Academic Press. p. 217-246.
146. Mullegama, S.V., et al., *RAI1 Overexpression Promotes Altered Circadian Gene Expression and Dyssomnia in Potocki-Lupski Syndrome.* J Pediatr Genet, 2017. **6**(3): p. 155-164.
147. Potocki, L., et al., *Molecular mechanism for duplication 17p11.2- the homologous recombination reciprocal of the Smith-Magenis microdeletion.* Nat Genet, 2000. **24**(1): p. 84-7.
148. Falco, M., S. Amabile, and F. Acquaviva, *RAI1 gene mutations: mechanisms of Smith-Magenis syndrome.* The application of clinical genetics, 2017. **10**: p. 85-94.
149. Elsea, S.H. and S.R. Williams, *Smith-Magenis syndrome: haploinsufficiency of RAI1 results in altered gene regulation in neurological and metabolic pathways.* Expert Rev Mol Med, 2011. **13**: p. e14.
150. Girirajan, S., et al., *How much is too much? Phenotypic consequences of Rai1 overexpression in mice.* Eur J Hum Genet, 2008. **16**(8): p. 941-54.
151. Williams, S.R., et al., *Smith-Magenis syndrome results in disruption of CLOCK gene transcription and reveals an integral role for RAI1 in the maintenance of circadian rhythmicity.* Am J Hum Genet, 2012. **90**(6): p. 941-9.
152. Gréchez-Cassiau, A., et al., *The hepatic circadian clock regulates the choline kinase α gene through the BMAL1-REV-ERBα axis.* Chronobiol Int, 2015. **32**(6): p. 774-84.
153. Song, J., et al., *Polymorphism of the PEMT gene and susceptibility to nonalcoholic fatty liver disease (NAFLD).* Faseb j, 2005. **19**(10): p. 1266-71.
154. Gorné, L.D., et al., *The mouse liver displays daily rhythms in the metabolism of phospholipids and in the activity of lipid synthesizing enzymes.* Chronobiol Int, 2015. **32**(1): p. 11-26.
155. Hua, X., et al., *Structure of the human gene encoding sterol regulatory element binding protein-1 (SREBF1) and localization of SREBF1 and SREBF2 to chromosomes 17p11.2 and 22q13.* Genomics, 1995. **25**(3): p. 667-73.
156. Shimano, H., et al., *Sterol regulatory element-binding protein-1 as a key transcription factor for nutritional induction of lipogenic enzyme genes.* J Biol Chem, 1999. **274**(50): p. 35832-9.
157. Lech, K., et al., *Dissecting Daily and Circadian Expression Rhythms of Clock-Controlled Genes in Human Blood.* J Biol Rhythms, 2016. **31**(1): p. 68-81.
158. Smith, A.C., et al., *Hypercholesterolemia in children with Smith-Magenis syndrome: del (17) (p11.2p11.2).* Genet Med, 2002. **4**(3): p. 118-25.
159. Pereira, C.F., et al., *ESCs require PRC2 to direct the successful reprogramming of differentiated cells toward pluripotency.* Cell stem cell, 2010. **6**(6): p. 547-556.
160. Klosin, A., et al., *Transgenerational transmission of environmental information in C. elegans.* Science, 2017. **356**(6335): p. 320-323.
161. Arnardottir, E.S., et al., *Blood-gene expression reveals reduced circadian rhythmicity in individuals resistant to sleep deprivation.* Sleep, 2014. **37**(10): p. 1589-1600.
162. Luco, R.F., et al., *Regulation of alternative splicing by histone modifications.* Science, 2010. **327**(5968): p. 996-1000.
163. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic acids research, 2002. **30**(1): p. 207-210.
164. Toolkit, N.S.; Available from: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software.
165. Kaneko, S., et al., *PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells.* Nat Struct Mol Biol, 2013. **20**(11): p. 1258-64.
166. Fang, D., et al., *H3.3K27M mutant proteins reprogram epigenome by sequestering the PRC2 complex to poised enhancers.* Elife, 2018. **7**.

167. Gu, Z., et al., *Loss of EZH2 Reprograms BCAA Metabolism to Drive Leukemic Transformation.* Cancer Discov, 2019. **9**(9): p. 1228-1247.
168. Liu, Y., et al., *Convergence of oncogenic cooperation at single-cell and single-gene levels drives leukemic transformation.* Nat Commun, 2021. **12**(1): p. 6323.
169. Andrews, S., *FastQC: a quality control tool for high throughput sequence data*. 2010, Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.
170. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-2120.
171. Dobin, A. *STAR manual 2.7.0a*. 2019.
172. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature methods, 2012. **9**(4): p. 357-359.
173. Li, H., et al., *The sequence alignment/map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.
174. Ramírez, F., et al., *deepTools2: a next generation web server for deep-sequencing data analysis.* Nucleic acids research, 2016. **44**(W1): p. W160-W165.
175. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome biology, 2014. **15**(12): p. 1-21.
176. Stark R, B.G. *DiffBind: differential binding analysis of ChIP-Seq peak* 2011; Available from: http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pd.
177. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
178. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome biology, 2008. **9**(9): p. 1-9.
179. Zang, C., et al., *A clustering approach for identification of enriched domains from histone modification ChIP-Seq data.* Bioinformatics, 2009. **25**(15): p. 1952-1958.
180. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.
181. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57.
182. Putri, G.H., et al., *Analysing high-throughput sequencing data in Python with HTSeq 2.0.* arXiv preprint arXiv:2112.00939, 2021.
183. McLean, C.Y., et al., *GREAT improves functional interpretation of cis-regulatory regions.* Nature biotechnology, 2010. **28**(5): p. 495-501.
184. Wickham, H., *ggplot2: elegant graphics for data analysis*. 2016: springer.
185. Blighe K, R.S., Lewis M, *EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling.* R package version 1.12.0, 2021. **https://github.com/kevinblighe/EnhancedVolcano**.
186. Robinson, J.T., et al., *Integrative genomics viewer.* Nature biotechnology, 2011. **29**(1): p. 24-26.
187. Durinck, S., et al., *Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.* Nature protocols, 2009. **4**(8): p. 1184-1191.
188. Karolchik, D., et al., *The UCSC genome browser database.* Nucleic acids research, 2003. **31**(1): p. 51-54.
189. Dabbish, L., et al. *Social coding in GitHub: transparency and collaboration in an open software repository*. in *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 2012.
190. Deaton, A.M. and A. Bird, *CpG islands and the regulation of transcription.* Genes Dev, 2011. **25**(10): p. 1010-22.
191. Wu, H., et al., *Structure of the catalytic domain of EZH2 reveals conformational plasticity in cofactor and substrate binding sites and explains oncogenic mutations.* PLoS One, 2013. **8**(12): p. e83737.
192. Long, Y., et al., *Conserved RNA-binding specificity of polycomb repressive complex 2 is achieved by dispersed amino acid patches in EZH2.* Elife, 2017. **6**.

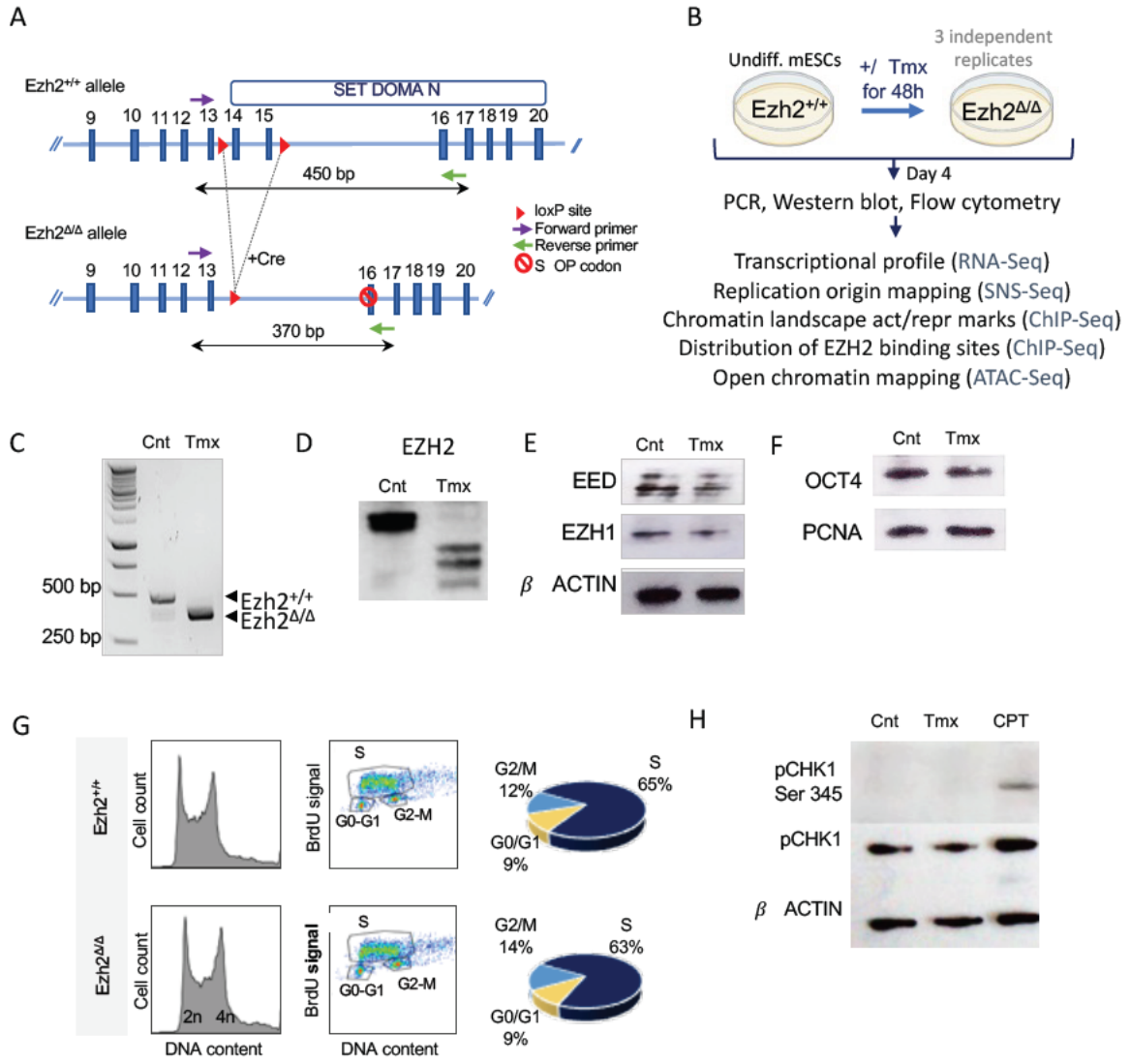
193. Gentile, C. and M. Kmita, *Polycomb Repressive Complexes in Hox Gene Regulation: Silencing and Beyond: The Functional Dynamics of Polycomb Repressive Complexes in Hox Gene Regulation.* Bioessays, 2020. **42**(10): p. e1900249.
194. Holland, P.W.H., H.A.F. Booth, and E.A. Bruford, *Classification and nomenclature of all human homeobox genes.* BMC biology, 2007. **5**: p. 47-47.
195. Danis, E., et al., *Specification of a DNA replication origin by a transcription complex.* Nat Cell Biol, 2004. **6**(8): p. 721-30.
196. Tie, F., et al., *CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing.* Development, 2009. **136**(18): p. 3131-41.
197. Lavarone, E., C.M. Barbieri, and D. Pasini, *Dissecting the role of H3K27 acetylation and methylation in PRC2 mediated control of cellular identity.* Nature Communications, 2019. **10**(1): p. 1679.
198. Pasini, D., et al., *Characterization of an antagonistic switch between histone H3 lysine 27 methylation and acetylation in the transcriptional regulation of Polycomb group target genes.* Nucleic Acids Res, 2010. **38**(15): p. 4958-69.
199. Li, Y., et al., *Integrated bioinformatics analysis of chromatin regulator EZH2 in regulating mRNA and lncRNA expression by ChIP sequencing and RNA sequencing.* Oncotarget, 2016. **7**(49): p. 81715-81726.
200. Popovic, R., et al., *Histone methyltransferase MMSET/NSD2 alters EZH2 binding and reprograms the myeloma epigenome through global and focal changes in H3K36 and H3K27 methylation.* PLoS Genet, 2014. **10**(9): p. e1004566.
201. Bell, S.P., R. Kobayashi, and B. Stillman, *Yeast origin recognition complex functions in transcription silencing and DNA replication.* Science, 1993. **262**(5141): p. 1844-1849.
202. Hou, Y., et al., *Novel Roles of G-quadruplexes on Enhancers in human chromatin*. 2021.
203. Wells, J., et al., *Target Gene Specificity of E2F and Pocket Protein Family Members in Living Cells.* Molecular and cellular biology, 2000. **20**: p. 5797-807.
204. Miyake, Y., et al., *Novel splicing variant of mouse Orc1 is deficient in nuclear translocation and resistant for proteasome-mediated degradation.* J Biol Chem, 2005. **280**(13): p. 12643-52.
205. Spiegel, J., S. Adhikari, and S. Balasubramanian, *The Structure and Function of DNA G-Quadruplexes.* Trends in Chemistry, 2020. **2**(2): p. 123-136.
206. Egan, B., et al., *An Alternative Approach to ChIP-Seq Normalization Enables Detection of Genome-Wide Changes in Histone H3 Lysine 27 Trimethylation upon EZH2 Inhibition.* PloS one, 2016. **11**(11): p. e0166438-e0166438.
207. Blanco, E., L. Di Croce, and S. Aranda, *SpikChIP: a novel computational methodology to compare multiple ChIP-seq using spike-in chromatin.* NAR Genomics and Bioinformatics, 2021. **3**(3).
208. Ura, H., et al., *STAT3 and Oct-3/4 control histone modification through induction of Eed in embryonic stem cells.* J Biol Chem, 2008. **283**(15): p. 9713-23.
209. Helmrich, A., M. Ballarino, and L. Tora, *Collisions between Replication and Transcription Complexes Cause Common Fragile Site Instability at the Longest Human Genes.* Molecular Cell, 2011. **44**(6): p. 966-977.
210. Lenstra, T.L., et al., *Transcription Dynamics in Living Cells.* Annual review of biophysics, 2016. **45**: p. 25-47.
211. Bartman, C.R., et al., *Transcriptional Burst Initiation and Polymerase Pause Release Are Key Control Points of Transcriptional Regulation.* Molecular Cell, 2019. **73**(3): p. 519-532.e4.
212. Okano-Uchida, T., et al., *Endoreduplication of the mouse genome in the absence of ORC1.* Genes & development, 2018. **32**(13-14): p. 978-990.

SUPPLEMENTARY

"CHROMATIN LANDSCAPE, TRANSCRIPTION AND DNA REPLICATION ORIGIN POSITIONING"

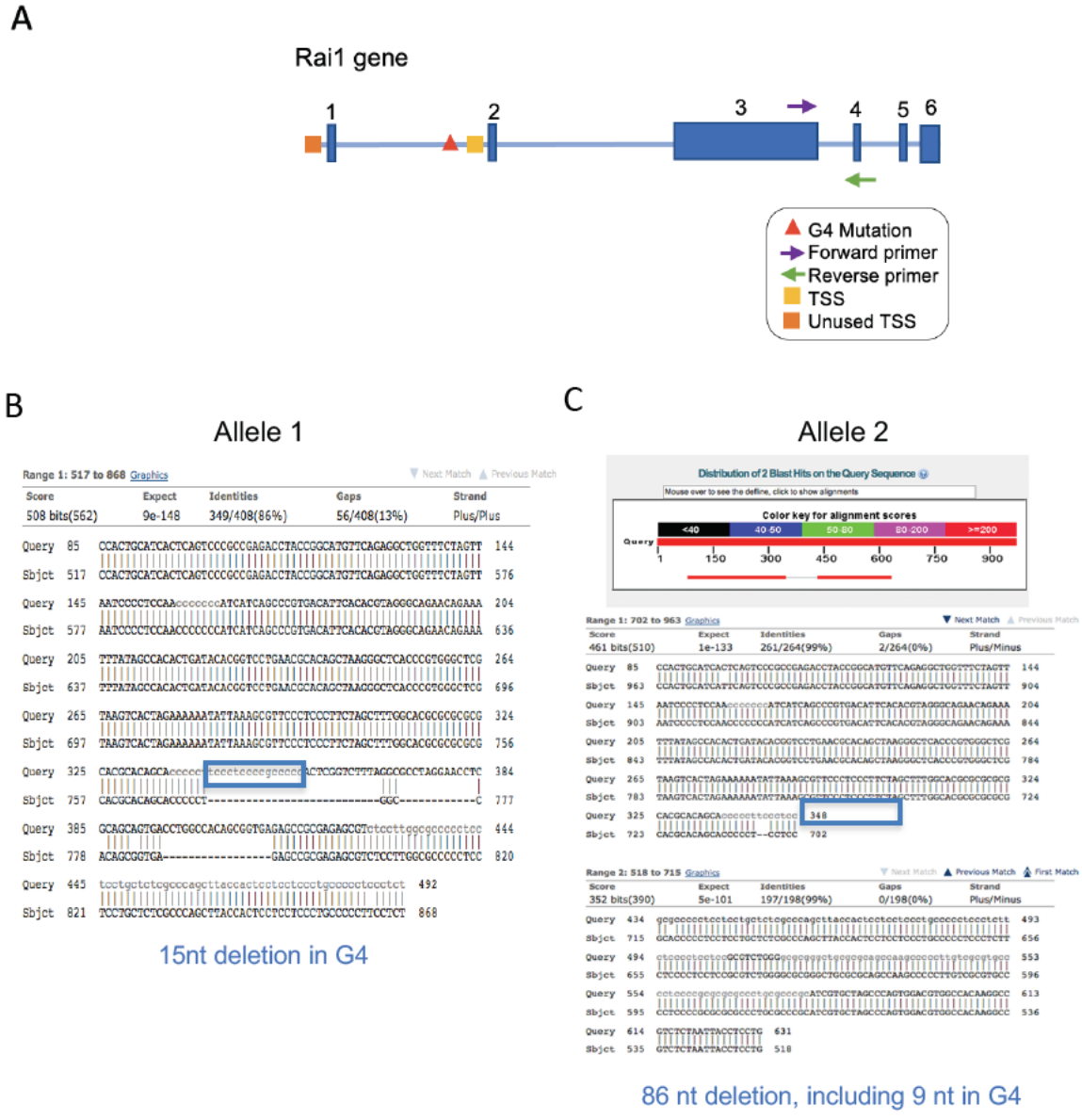
NEREA MURUGARREN GARRIDO

ID

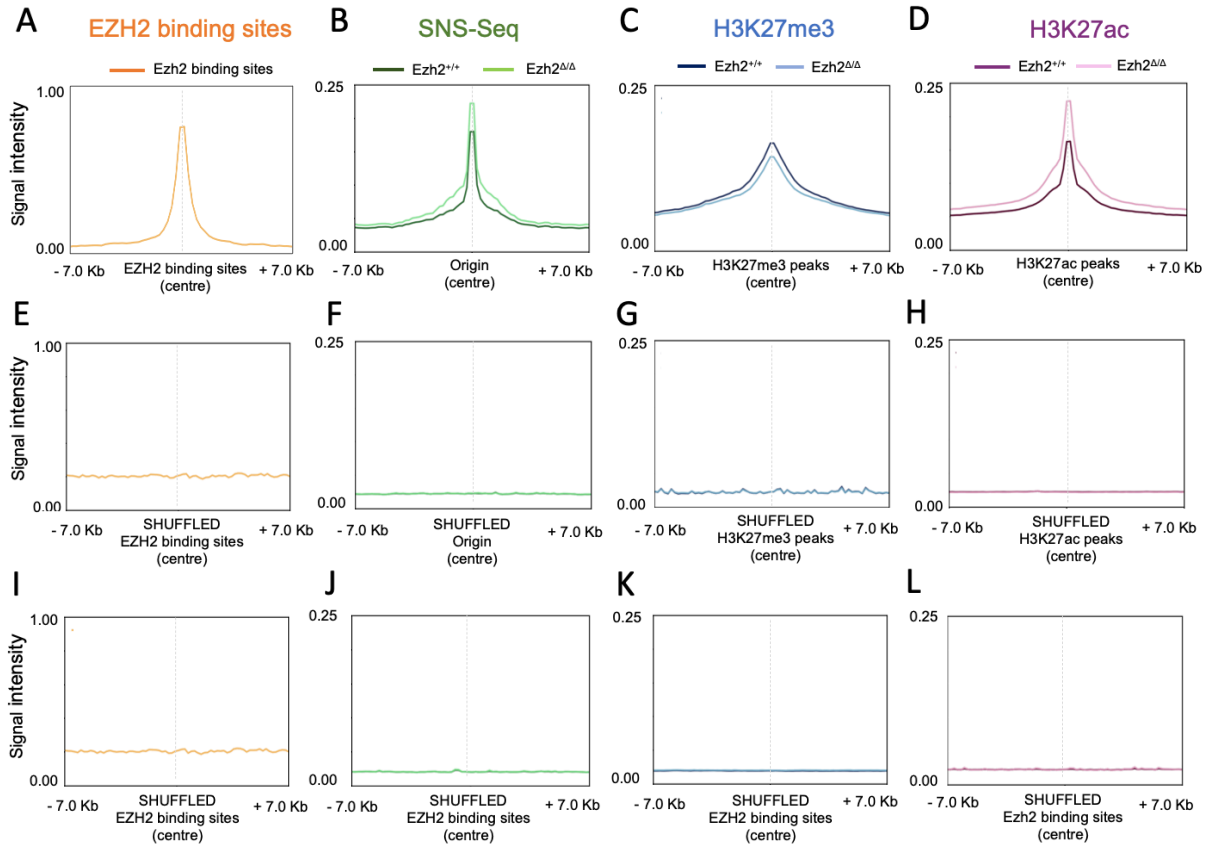


Supplementary Figure 1. Ezh2$^{\Delta/\Delta}$ cell line genomic mutation and wet-lab control experiments. (A) Schemat c representat on of tamox fen- nduc b e Ezh2$^{\Delta/\Delta}$ des gn us ng the Cre/LoxP system. Exons 14 and 15 of EZH2 SET doma n f anked by LoxP s tes. The express on of Cre recomb nase by add ng tamox fen (Tam) eads to the exc s on of the LoxP f anked reg ons, eav ng on y one LoxP recomb nase recogn t on s te n the mESC genome. PCR forward and reverse pr mers ocat on a so shown. **(B)** Schemat c of the overa exper menta workf ow used for an ntegrated rep cat on or g n, transcr pt on and ep genet c mapp ng approach n Ezh2$^{\Delta/\Delta}$ cond t ons. Step 1: Ezh2$^{\Delta/\Delta}$ mouse ES ce s are cu tured, then nduct on n t ated by treatment w th tamox fen (Tam). n = 3 ndependent ce cu ture rep cates were performed for each cond t on. Step 2: Character st cs of knockout n treated and contro Ezh2$^{\Delta/\Delta}$ mESC ce s were determ ned by western b ot, PCR, and ce cyc e ana ys s performed us ng f ow cytometry. Step 3: To ga n ns ghts nto the nature of the re at onsh ps between or g n act v ty and nearby transcr pt on, Ezh2$^{+/+}$ and Ezh2$^{\Delta/\Delta}$ mESC ce s were subjected to SNS-Seq, RNA-Seq and ChIP-Seq exper ments (aga nst H3K27me3 and H3K27ac marks) from the same batch of ce s. Step 4: NGS data ana ys s was performed f rst w th mapp ng of reads to the genome and peak ca ng and then compar ng / corre at ng the peaks from these three ana yt ca approaches. **(C)** PCR ana yzes of genom c DNA samp es generated n Ezh2$^{+/+}$ and Ezh2$^{\Delta/\Delta}$ mESCs to va date SET doma n exc s on at genom c eve . **(D-F)** Western b ot ana yzes of Ezh2$^{+/+}$ and Ezh2$^{\Delta/\Delta}$ ce s tota prote n extracts w th the nd cated ant bod es (ant - EZH2, EED, EZH1, β-ACTIN, OCT4, PCNA). OCT4: Octamer-b nd ng

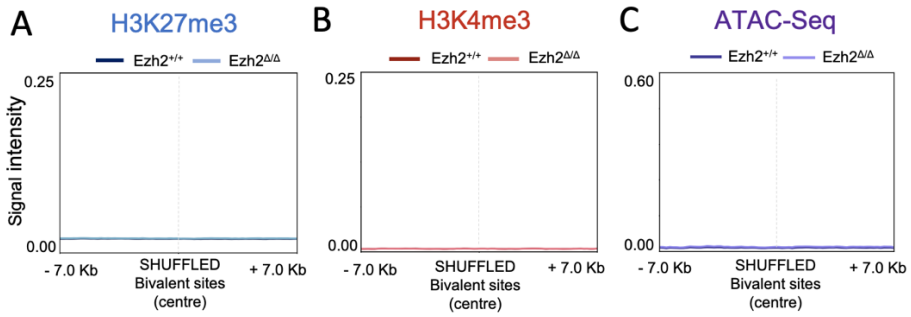
transcr pt on factor 4; PCNA: Pro ferat ng ce nuc ear ant gen. **G** Ce cyc e d str but on of Ezh2$^{+/+}$ and Ezh2$^{\Delta/\Delta}$ mESCs determ ned by f ow cytometry after abe ng of S-phase ce s w th the thym d ne ana og BrdU. The fract on of ce s n the d fferent ce cyc e phases s nd cated. **H** Western b ot ana yzes of Ezh2$^{+/+}$ and Ezh2$^{\Delta/\Delta}$ ce s tota prote n extracts w th the nd cated ant bod es (ant - phosphory ated CHK1, CHK1, β-ACTIN). Cnt: Contro ; Tmx: Tamox fen treatment; CPT: Camptothec n treatment (topo somerase 2 nh b tor); CHK1: Checkpo nt k nase 1.



Supplementary Figure 2. 3T3$^{\Delta G4}$ cell line genomic mutation. (A) Mouse Ra 1 gene w th exons abe ed 1-6 show ng the ocat on of the G4 mutat on (3T3$^{\Delta G4}$ ce s). The Ra 1 transcr pt can be obta ned from two TSS s tes, wh ch are marked n the f gure. On y one of the two TSS s tes s expressed, thus on y the express on of the TSS s te marked as ye ow s nvest gated n th s thes s. PCR forward and reverse pr mers ocat on a so shown. **(B)** Mutated a e e 1 n 3T3$^{\Delta G4}$ ce s. **(C)** Mutated a e e 2 n 3T3$^{\Delta G4}$ ce s.



Supplementary Figure 3. **Control read pile up plots for M. Méchali Lab and EZH2 binding sites datasets.** Control read pile up plots displaying the EZH2 binding sites, SNS-Seq, H3K27me3 and H3K27ac (from left to right) signal profile from $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ mESC cells centered on the regions displayed (± 7 Kb). Shuffled regions refer to randomized data. EZH2 binding sites sources can be found on Table 2.



Supplementary Figure 4. **Control read pile up plots for Z. Gu et al. datasets. (A-C)** Control read pile up plots displaying the H3K27me3, H3K4me3 and ATAC-Seq signal profile from $Ezh2^{+/+}$ and $Ezh2^{\Delta/\Delta}$ mutant potent cells centered on randomized Bivalent domain regions (shuffled regions, ± 7 Kb). Datasets downloaded from Z. Gu et al., performed with the use of LSK murine hematopoietic stem cells [167].

Dataset type	Phenotype	Number of peaks	% from All peaks
SNS-Seq	**All**	**55,553**	**100%**
SNS-Seq	$Ezh2^{+/+}$	25,357	45.64%
SNS-Seq	$Ezh2^{\Delta/\Delta}$	50,684	91.23%
SNS-Seq	Upregu ated n $Ezh2^{\Delta/\Delta}$	4,839	8.71%
SNS-Seq	Downregu ated n $Ezh2^{\Delta/\Delta}$	113	0.20%
SNS-Seq	Unchanged n $Ezh2^{\Delta/\Delta}$	23,776	42.80%
ChIP-Seq H3K27me3	**All**	**15,374**	**100%**
ChIP-Seq H3K27me3	$Ezh2^{+/+}$	13,881	90.28%
ChIP-Seq H3K27me3	$Ezh2^{\Delta/\Delta}$	13,725	89.27%
ChIP-Seq H3K27me3	Upregu ated n $Ezh2^{\Delta/\Delta}$	39	0.25%
ChIP-Seq H3K27me3	Downregu ated n $Ezh2^{\Delta/\Delta}$	419	2.76%
ChIP-Seq H3K27ac	**All**	**25,662**	**100%**
ChIP-Seq H3K27ac	$Ezh2^{+/+}$	24,553	95.67%
ChIP-Seq H3K27ac	$Ezh2^{\Delta/\Delta}$	15,971	62.23%
ChIP-Seq H3K27ac	Upregu ated n $Ezh2^{\Delta/\Delta}$	7,157	27.88%
ChIP-Seq H3K27ac	Downregu ated n $Ezh2^{\Delta/\Delta}$	113	0.04%
ChIP-Seq EZH2 binding sites (BS)	**EZH2 binding sites (BS)**	**10,114**	**100%**
ChIP-Seq Bivalent peaks (H3K4me3 & H3K27me3)	**Bivalent peaks (H3K27me3 & H3K4me3)**	**16,059**	**100%**
RNA-Seq	**All**	**10,133**	**100%**
RNA-Seq	Upregu ated n $Ezh2^{\Delta/\Delta}$	524	5.17%
RNA-Seq	Downregu ated n $Ezh2^{\Delta/\Delta}$	526	5.19%

Supplementary Table 1. Number of peaks for each dataset. Tab e show ng the number of peaks for each SNS-Seq, ChIP-Seq and RNA-Seq datasets. D fferences w th *p* va ues or FDR (Fa se d scovery rate) ≤ 0.05 were cons dered as stat st ca y s gn f cant to obta n upregu ated and downregu ates peaks.