

The genome-wide analysis of the specification and activation of  
human chromosomal DNA replication origin sites

by

Helen Sarah Wilkes

St Catharine's College, Cambridge



December 2020

This dissertation is submitted for the degree of Doctor of Philosophy at  
the University of Cambridge, England.

“As I am educated, I learn more and more about less and less,  
until I know a heck of a lot about bugger all.”

*pp. Adrian Dunn*

## **Preface**

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This thesis is not substantially the same as any that I have submitted, or is being concurrently submitted for a degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This thesis does not exceed the word limit of 60,000 words as stipulated by the Regulations of the Doctor of Philosophy examination for Biological Sciences.

# **The genome-wide analysis of the specification and activation of human DNA replication origin sites**

*Helen Sarah Wilkes*

## **Abstract:**

### *Background:*

Human DNA replication is a critical cellular process that initiates at tens of thousands of DNA replication origins. Next-generation sequencing (NGS) enabled the elucidation of some characteristics that define human replication origins.

Initiation-site sequencing (iniSeq) utilises the established human cell-free system for *in vitro* DNA replication, whereby a digoxigenin tag is introduced into newly synthesised DNA. Newly synthesised DNA was separated from total genomic DNA by immunoprecipitation and subjected to NGS.

Additionally, the replication factors Y RNAs and *Xenopus laevis* Nucleosome Remodelling and histone Deacetylation (xNuRD) complex are essential for DNA replication initiation, but the mechanism(s) involved remain unelucidated. The precise mechanisms of DNA replication, and factors involved in human replication origin specification and activation are not fully known.

### *My work:*

I developed and improved upon iniSeq to produce a novel DNA replication origin NGS method. Density-substitution initiation-site sequencing (ds-iniSeq) exploited the semi-conservative nature of replicated DNA to identify human replication origins and determine their relative activities. During a short *in vitro* replication reaction, newly synthesised DNA incorporated a heavy nucleotide (BrdUTP) which resulted in Heavy-Light replicated and Light-Light unreplicated DNA. This density substitution enabled reliable separation of replicated and unreplicated DNA, which underwent NGS and processing by MACS peak calling to identify the origins; Heavy-Light DNA: Light-Light DNA ratios at identified replication origins determined their relative activities.

I identified ~14,000 discrete replication origins and showed that CpG islands and CpG island-promoters were dominant features associated with high origin activity. Origins identified by ds-iniSeq (ds-iniSeq origins) strongly colocalised with transcription start sites and active genes, but transcriptional activity did not correlate with origin activity. Ds-iniSeq origins colocalised with previously identified early firing origin-associated histone marks. I employed ds-iniSeq to examine how Y RNAs and xNuRD impact replication origin specification and activation

genome-wide, by manipulating the *in vitro* replication reactions. Y RNA removal resulted in fewer ds-iniSeq origins with lower activities genome-wide. xNuRD restored the number and activities of origins genome-wide to similar levels when in the presence of Y RNAs.

I adapted ds-iniSeq to assess replication elongation; density-substitution elongation-site sequencing (ds-eloSeq) had a longer replication reaction incubation. Conducting ds-eloSeq alongside ds-iniSeq enabled the assessment of the transition from replication initiation to elongation, and the impact of replication factors on this relationship. I performed a preliminary analysis of the impact of Y RNA removal and xNuRD addition on elongation which revealed reduced site activities of DNA regions adjacent to the corresponding ds-iniSeq origins in the absence of Y RNAs. I have also endeavoured to develop a new method for calling replicating ds-eloSeq sites as MACS was inappropriate for these samples.

Finally, Y RNAs bind to various proteins during DNA replication. From mass spectrometric data (generated by M.Kowalski), I identified the Polycomb repressive complex 2 (PRC2) as a Y RNA binding complex that possessed DNA replication activity (in the human cell-free system). I showed that PRC2 plays a stimulatory role in DNA replication initiation.

In summary, I have established a novel NGS method for origin identification that determines relative replication origin activity and used it to assess the impact of genomic and epigenetic features, and hY RNAs and xNuRD, on origin specification and activation. I have further developed a NGS method to examine elongation and, using biochemical techniques, have characterised PRC2 as a stimulatory DNA replication initiation factor.

## **Acknowledgements**

I would like to express my sincerest gratitude my collaborators Professor Sir Jim Smith and his group at the Francis Crick Institute and Dr Julien Sale and his group at the LMB-MRC, without whom, this PhD project would not be possible. I would also like to thank St Catharine's college, Cambridge for their financial support.

Furthermore, I would like to thank my Lab group for their support and contribution throughout my PhD and my advisors, Dr Tim Weil and Dr Matthias Landgraf for their continued feedback and sound advice. Moreover, I would like to thank my friend Dr Souradip Mookerjee for his help with my bioinformatical analysis and my former college tutor, the late Dr Phillip Oliver, for his support and contribution to ensuring I had the financial support to carry out my PhD. Additionally, I would like to thank the Zoology department, where I carried out my PhD, and my departmental colleagues, particularly Dr Alice Rees for her feedback on my thesis, our cleaner Zoe for entertaining conversations and our tearoom assistant, Nicola who ensured the continual supply of tea.

I would like to express my deepest appreciation to my mother, Dr Elaine McCash for everything she has done; from her unwavering encouragement and support to her continuing financial contributions, guidance and exquisite proof reading skills. Finally, I would like to thank my family, particularly my little brother who, despite having no desire to learn about biology, let me explain my work to him anyway, my pets Sophie and the late Nutmeg, and my brother's dog Tilly, for their company when working from home and writing up, and my stepfather who always provides a unique perspective, and my close friends for their unrelenting support which has just about kept me sane.

## Abbreviations

AEBP2	– Adipocyte Enhancer-Binding Protein 2
ASF1	– Anti-silencing function 1
CDC 6/45	– Cell division cycle protein 6/ 45
CDK	– Cyclin-dependent kinase
CDT1	– Chromatin licensing and DNA replication factor 1
CGI	– CpG island
CGI-promoter	– CpG island promoter
CHD 3/4	– Chromodomain Helicase DNA binding protein 3/ 4
Control (ds-iniSeq) origins	– ds-iniSeq origins identified in the control condition
CTCF	– CCCTC-binding factor
DDK	– Dbf4-dependent kinase
dig-dUTP	– Digoxigenin d-UTP
EED	– Embryonic Ectoderm Development
EFOA	– Early Firing Origin-Associated
ENCODE	– Encyclopaedia of DNA elements
EZH1/2	– Enhancer of Zeste Homologue 1/2
FACT	– Facilitates chromatin transcription
G4	– G-quadruplex
GWAS	– Genome-wide association study
h/xY RNA	– Human/ <i>Xenopus laevis</i> Y RNA
HDAC m/1/2	– Histone deacetylase protein m/1/ 2
HL DNA	– Heavy-Light DNA
Human cytosol	– Cytosolic extract from proliferating human cells
ID	– Immunodepletion

iniSeq	– Initiation site sequencing
IR	– Interquartile range
LAD	– Lamina Associated Domain
LFOA	– Late Firing Origin-Associated
LL DNA	– Light-Light DNA
MACS	– Model-based analysis of ChIP-seq
MBD 2/3	– Methyl-CpG binding protein 2/ 3
MBT	– Mid-blastula transition
MCM 2-7/10	– Mini-chromosome maintenance proteins 2-7/ 10
Mock-depleted (ds-iniSeq) origins	– ds-iniSeq origins identified in the mock-depleted condition
MTA 1/2/3	– Metastasis- associated protein 1/ 2/ 3
ncRNA	– Non-coding RNA
NGS	– Next generation sequencing
nHL-nLL	– HL DNA read count normalised to total read count of the file minus LL DNA read count normalised to total read count of the file
OK-seq	– Okazaki sequencing
ORC	– Origin recognition complex
PCNA	– Proliferating cell nuclear antigen
PRC1/2	– Polycomb repressive complex ½
pre-RC	– Pre-replication complex
ds-iniSeq origins	– Replicate 1 origins overlapping origins present in replicates 2 and 3
ds-iniSeq RS	– Randomised sites corresponding to the ds-iniSeq origins
ds-iniSeq23 origins	– Replicate 2 origins overlapping origins in replicate 3 but not replicate 1

ds-iniSeq23 RS	– Randomised sites corresponding to the ds-iniSeq23 origins
RbAp 46/48	– Retinoblastoma-binding protein 46/ 48
RI	– Refractive index
RNP	– Ribonucleoprotein
RPA	– Replication protein A
SNS-seq	– Small nascent strand sequencing
SUZ12	– Suppressor of Zeste 12
TAD	– Topologically Associated Domain
TSS	– Transcription start site
xNuRD	– <i>Xenopus laevis</i> Nucleosome Remodelling and histone Deacetylase complex
xNuRD addition (ds-iniSeq) origins	– ds-iniSeq origins identified in the xNuRD addition condition
Y RNA-depleted (ds-iniSeq) origins	– ds-iniSeq origins identified in the Y RNA-depleted condition

# Contents

<b>Preface</b> .....	<b>I</b>
<b>Abstract</b> .....	<b>II</b>
<b>Acknowledgements</b> .....	<b>IV</b>
<b>Abbreviations</b> .....	<b>V</b>
<b>Contents</b> .....	<b>VIII</b>
<b>Chapter 1: Introduction</b> .....	<b>1</b>
1.1 The Cell Cycle and Eukaryotic Chromosomal DNA Replication.....	1
1.2 Eukaryotic Chromosomal DNA Replication Initiation.....	6
1.3 Eukaryotic Chromosomal DNA Replication Origins.....	11
1.3.1 Global nuclear architecture.....	14
1.3.2 Primary sequence.....	16
1.3.3 Epigenetics.....	19
1.4 Assessment of DNA Replication on the Human Cell-Free System.....	23
1.5 Initiation-Site Sequencing (iniSeq).....	26
1.6 Y RNAs and their Role in Chromosomal DNA Replication.....	28
1.7 Early <i>Xenopus laevis</i> Embryo DNA replication and <i>X. laevis</i> Nucleosome Remodelling and Histone Deacetylation (xNuRD) complex.....	35
<b>Chapter 2: Aims and Objectives</b> .....	<b>39</b>
<b>Chapter 3: Materials and Methods</b> .....	<b>41</b>
3.1 Cell culture.....	41
3.2 Cell Propagation.....	41
3.3 Preparation of Human cytosolic extract, S100.....	41
3.4 Preparation of EJ30 late G1- and S- phase template nuclei.....	41

3.4.1 Late G1-phase synchronisation.....	41
3.4.2 S-phase synchronisation.....	42
3.4.3 Nuclei isolation preparation.....	42
3.5 <i>In vitro</i> DNA replication experiment using the Human cell-free system.....	42
3.5.1 Preparation of the replication reactions.....	42
3.5.2 Manipulation of the replication reactions.....	43
3.5.3 Stopping the replication reaction and slide preparation and analysis....	44
3.6 Preparation of xNuRD from <i>Xenopus laevis</i> eggs.....	45
3.6.1 Ammonium sulphate precipitation.....	45
3.6.2 Sucrose gradient ultracentrifugation.....	45
3.6.3 Bradford (BioRad) assay.....	45
3.7 Density-Substitution Initiation-site sequencing (ds-iniSeq).....	46
3.7.1 Replication reactions to generate replicated DNA.....	46
3.7.2 Isolation of total genomic DNA.....	46
3.7.3 Separation of HL and LL DNA.....	47
3.7.4 Preparation of HL and LL DNA for sequencing.....	47
3.7.5 Next-generation DNA sequencing.....	48
3.8 Density-Substitution elongation-site sequencing (ds-eloSeq).....	48
3.9 Bioinformatics.....	49
3.9.1 Processing and alignment of sequencing read – ds-iniSeq and ds-eloSeq.....	49
3.9.2 Peak calling – ds-iniSeq.....	49
3.9.3 Assessment of elongation – ds-iniSeq and ds-eloSeq.....	49
3.9.4 Peak calling – Wilkes-Mookerjee (WM) method.....	49
3.9.5 Bioinformatical analysis.....	50
3.9.6 Realignment of 3 <sup>rd</sup> party genomic data.....	51

3.10 Western blot.....	51
3.11 Immunodepletion of human cytosolic extract.....	52
3.12 Agarose gels.....	52
3.13 Total RNA preparation.....	52
3.14 Total RNA sequencing.....	53

**Chapter 4: The development of density-substitution initiation-site sequencing (ds-iniSeq)..... 54**

4.1 Introduction.....	54
4.2 Results and Discussion.....	55
4.2.1 The ds-iniSeq method.....	55
4.2.2 Preliminary ds-iniSeq experiment.....	58
4.2.3 Production of ds-iniSeq replicates.....	62
4.2.4 NGS sequencing.....	66
4.2.5 Origin calling by MACS.....	68
4.2.6 ds-iniSeq replicates – called origins and overlap analysis.....	87
4.2.7 Conclusions.....	95

**Chapter 5: The analysis of the standard density-substitution initiation-site sequencing (ds-iniSeq) experiment..... 97**

5.1 Introduction.....	97
5.2 Results and discussion.....	97
5.2.1 Origins per chromosome.....	97
5.2.2 Comparison to alternative origin identification methods.....	100
5.2.2a iniSeq overlap.....	100
5.2.2b SNS-seq overlap.....	102
5.2.2c OK-seq overlap.....	105
5.2.2d Bubble-seq overlap.....	107

5.2.3 Replication timing.....	110
5.2.4 Percentage GC content.....	113
5.2.5 Comparison to genomic features.....	115
5.2.5a Transcription start sites (TSS).....	121
5.2.5b CpG islands (CGIs).....	121
5.2.5c G-quadruplexes (G4s).....	122
5.2.5d Multiple overlap analyses.....	123
5.2.6 Polarity of G4s.....	124
5.2.7 Genes.....	126
5.2.8 Promoters and enhancers.....	131
5.2.9 Comparison to epigenetic features.....	136
5.2.10 Interplay between genomic and epigenetic features.....	144
5.3 Discussion.....	149

**Chapter 6: The role of hY RNAs and xNuRD for the activation of human DNA replication origins, as determined by ds-iniSeq.....**

<b>158</b>	<b>158</b>
6.1 Introduction.....	158
6.2 Results.....	161
6.2.1 Comparison of Control ds-iniSeq origins and Mock-depleted ds-iniSeq origins.....	161
6.2.2 Comparison of mock-depleted, Y RNA-depleted and xNuRD addition ds-iniSeq origins.....	165
6.2.3 How do Y RNAs and xNuRD impact DNA replication origins? – overlap and origin activities.....	169
6.2.4 How do Y RNAs and xNuRD impact DNA replication origins? – replication timing.....	175
6.2.5 How do Y RNAs and xNuRD impact DNA replication origins? – genomic and epigenetic features.....	179

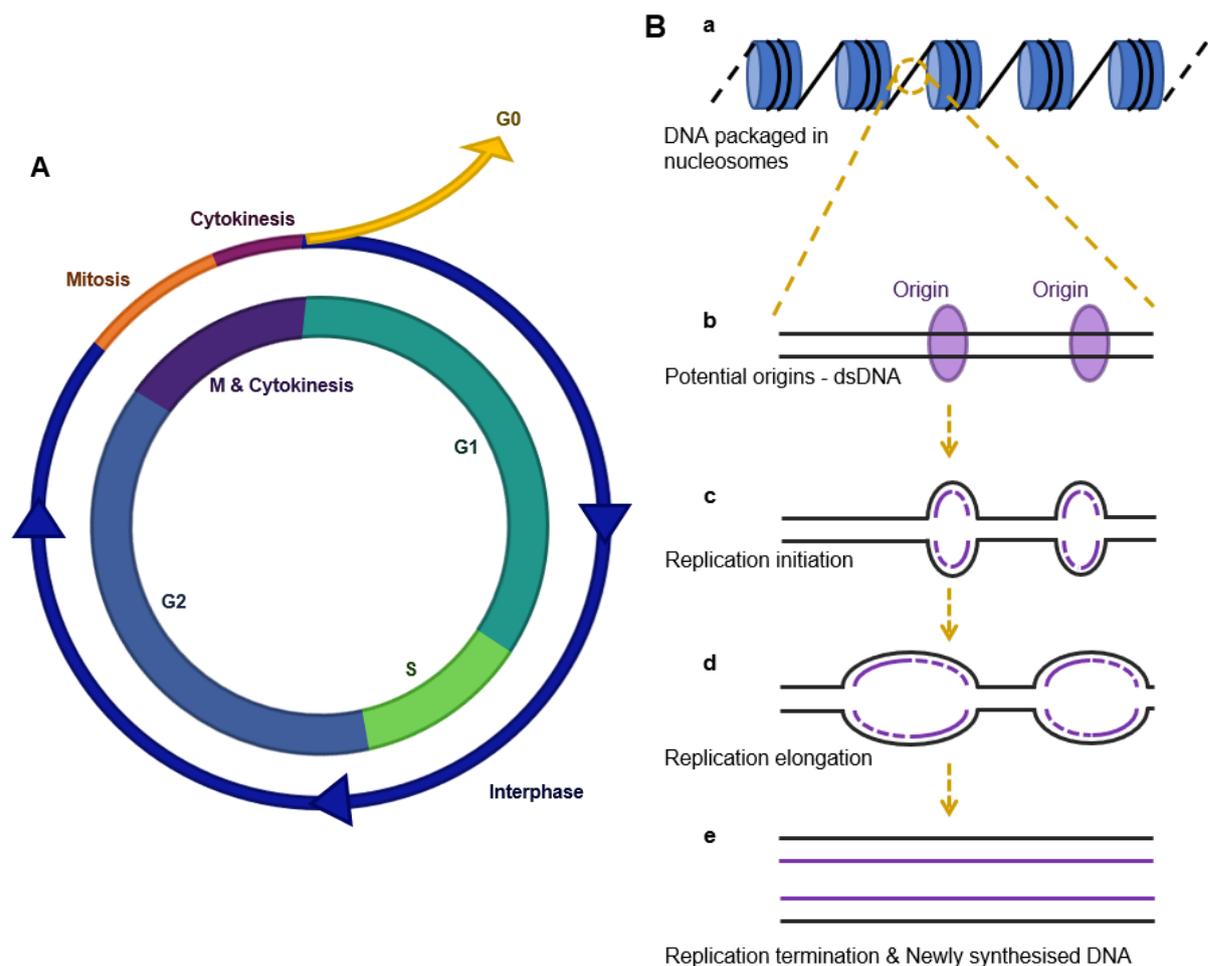
6.3 Discussion.....	189
<b>Chapter 7: The impact of Y RNAs and xNuRD on replication elongation using density-substitution elongation-site sequencing (ds-eloSeq).....</b>	<b>194</b>
7.1 Introduction.....	194
7.2 Results and discussion – Preliminary analysis of ds-eloSeq data.....	194
7.2.1 Preliminary analysis of elongation using ds-eloSeq data.....	198
7.2.2 Ds-iniSeq origin sites.....	199
7.2.3 Elongation away from ds-iniSeq origin sites.....	102
7.3 Results and discussion – Development of origin calling method in ds-eloSeq data.....	208
7.3.1 “Wilkes-Mookerjee (WM) method for origin calling in 3-hour ds-eloSeq data.....	208
7.3.2 WM origin calling in 15-minute ds-iniSeq samples.....	211
7.3.3 WM origin calling in 3-hour ds-eloSeq samples.....	213
<b>Chapter 8: A role for the chromatin remodelling complex, Polycomb Repressive Complex 2 (PRC2), in human DNA replication initiation.....</b>	<b>220</b>
8.1 Previous work.....	220
8.2 Polycomb-repressive complex 2 – identification and introduction.....	222
8.2.1 Identification.....	222
8.2.2 Introduction – PRC2 background.....	223
8.3 Results.....	230
8.3.1 This project.....	230
8.3.2 Antibody treatment of the cytosol.....	231
8.3.3 Chemical inhibitor treatment of the cytosol.....	234
8.3.4 Antibody and chemical inhibitor treatment of the cytosol.....	236

8.3.5 Immunodepletion of SUZ12 from the cytosol.....	238
8.3.6 Inhibition of PRC2 in S-phase template nuclei.....	242
8.4 Discussion.....	244
8.4.1 PRC2 is a stimulatory replication initiation factor.....	244
8.4.2 PRC2 and other chromatin remodellers in DNA replication.....	245
8.4.3 PRC2 and RNA binding.....	246
8.4.4 PRC2 and ds-iniSeq origins.....	247
8.4.5 Summary.....	249
<b>Chapter 9: Conclusion.....</b>	<b>250</b>
<b>References.....</b>	<b>252</b>
<b>Appendices.....</b>	<b>305</b>
Appendix – A3.....	305
Appendix – A4.....	310
Appendix – A5.....	318

# Chapter 1: Introduction

## 1.1 The Cell Cycle and Eukaryotic Chromosomal DNA Replication

The duplication of one parent cell to produce two genetically identical daughter cells is essential for the longevity of multi-cellular organisms. The cell cycle (Fig1.1A) is the series of events that result in the generation of these daughter cells. There are three key phases of the cell cycle: interphase, mitosis and cytokinesis (1).



**Figure 1.1:** (A) The schematic summary of the cell cycle where interphase consists of Growth phases 1 and 2 (G1 and G2), and the synthesis phase (S). Mitosis and cytokinesis follows interphase and once mitosis is complete the cell may continue through the cell cycle or enter quiescence (G0). (B) S-phase comprises semi-conservative DNA synthesis/replication through the following steps: (a) DNA is packaged into nucleosomes (blue cylinder) and contains potential DNA replication origin sites (pale purple) (b); (c) DNA replication begins with initiation where licensed origins are activated/fired and bidirectional replication forks are formed (parental DNA in black, and newly synthesised DNA in purple); (d) DNA replication moves on to the elongation stage, where the established replication forks extend into replication bubble; finally DNA replication enters the termination stage where adjacent replication bubbles merge and newly synthesised DNA is completed (e).

Mitosis and cytokinesis (M-phase) are the processes by which the single cell divides into two. Mitosis comprises four stages (prophase, metaphase, anaphase, and telophase) where the nuclear membrane is broken down; the chromosomes are condensed, aligned to the cell equator, and then pulled to the opposite poles of the cell by the centrioles and spindles. The nuclear membranes then reform around the genetic material and cytokinesis takes place, where the single cell undergoes cleavage to form two daughter cells. Following cytokinesis, the cells may continue through the cell cycle again or may diverge from the cell cycle and become quiescent (G<sub>0</sub>), whereby cell division no longer takes place (Fig1.1A) (1,2).

The cell cycle is predominantly spent in interphase, which consists of two growth stages (G<sub>1</sub> and G<sub>2</sub>-phases) either side of a synthesis stage (S-phase), during which DNA replication occurs, resulting in the whole genome being duplicated in preparation for cell division (1).

Accurate and regulated whole genome duplication is essential for the maintenance of genomic stability. The whole eukaryotic genome is replicated at approximately 1000 bases/minute/replication fork and this complete replication occurs only once per cell cycle (1,2). There is the potential for errors to occur during DNA replication. The estimated error rate of eukaryotic DNA polymerases is one in 10<sup>4</sup> – 10<sup>5</sup> nucleotides polymerised, which include base pair mismatches, insertions, and deletions (2–5). To ensure the fidelity of the duplicated genome, some DNA polymerases also possess a 3'-5' exonuclease activity and acts as a “proof-reading” mechanism. DNA proof-reading polymerases containing an exonuclease activity removes incorrect bases when they are recognised. Once removed, DNA replication can then continue (2,6,7). Proof-reading mechanisms reduces the spontaneous mutation/error rate to one in 10<sup>8</sup>-10<sup>10</sup> nucleotide polymerised (2,3,5,8).

Cell cycle checkpoints also work to ensure genome fidelity. There are three key checkpoints during the cell cycle where the cell monitors external signals and the internal environment prior to continuing with the cell cycle. The G<sub>1</sub>/S phase checkpoint triggers the cell committing to entering the cell cycle and assesses whether DNA is undamaged. If the genome is free from damage, the cell proceeds to the next phase. The G<sub>2</sub>/M checkpoint monitors cellular conditions prior to triggering mitosis. Notably, if DNA replication or DNA repair is incomplete, the cell cycle arrests until these are completed. Finally, the M checkpoint monitors the successful formation of spindles and their attachment to kinetochores; if these do not occur, mitosis is arrested (2,3,9,10).

Throughout S-phase, DNA replication is monitored for DNA damage, genomic stability, and replication fork progression. Should events that stall, or slow down DNA replication fork progression occur, the replication checkpoint mechanism is triggered (11,12). The replication checkpoint is a conserved kinase cascade that consists of three components: damage sensor proteins, signal transducers and effector proteins. The sensor proteins include ATM

(Ataxia Telangiectasia Mutated), ATR (ATM- and Rad3-related), DNA-PK (DNA-dependent protein kinase) and the Rad17-RFC and 9-1-1 complexes and recognise issues in replication and initiate the signal transducers such as Chk1 and Chk2 Ser/Thr kinases and Cdc25 phosphatases. In turn these transducers activate the effector proteins such as p53 and inactivate cyclin-dependent kinases, in order to promote cell cycle arrest until any issues can be resolved or initiate apoptosis when errors are irreparable (3,4,9,11–13).

On occasion, these checkpoints can fail, and DNA replication is carried out in an unregulated and inaccurate fashion, which can result in uncontrolled cell proliferation and ultimately lead to the development of diseases such as cancer (3,9).

A great deal of research has been conducted in order to elucidate the precise mechanisms by which the genome is faithfully replicated. In 1958, Meselson and Stahl used density substitution of nucleotides to establish that DNA was replicated in a semi-conservative fashion (Fig1.1B). This demonstrated that, of newly replicated double-stranded DNA (dsDNA), one strand comes from the original parent DNA, and the other, complementary strand, is newly synthesised using the parent strand as a template (14).

Eukaryotic DNA replication is a highly evolutionarily conserved process that is brought about by multiple and complex multi-protein machinery. There are three distinct stages that make up the DNA replication process: initiation, elongation and termination (1,15).

During the late G1-phase of the cell cycle, the pre-replication complex (pre-RC) forms at the tens of thousands of DNA replication origin sites throughout the genome (1,15), licensing the DNA for replication. When the cell cycle enters S-phase (which takes approximately 8 hours to complete in humans), protein kinases (CDK and DDK) can then activate DNA replication through a process known as origin firing (16–18).

During initiation, the DNA helix is unwound and unzipped by helicase and additional proteins are recruited to the pre-RC, thus leading to the formation of the pre-initiation complex. Once the pre-initiation complex, which includes DNA polymerases  $\alpha$ ,  $\delta$  and  $\epsilon$ , is present at the DNA, bidirectional DNA replication forks can be established (1,15,19,20).

DNA replication progresses to elongation when the established bidirectional replication forks transition into replication bubbles (Fig1.1B). RNA primers, approximately 10 nucleotides long, are laid down by DNA primase. These primers are used by the DNA polymerases to synthesise new DNA. DNA polymerase can only synthesise DNA in a 5' to 3' direction. This constraint, in conjunction with the antiparallel formation of the DNA double helix, results in asymmetric DNA replication, with new DNA is synthesised as part of the leading (running in the 5' to 3' direction) and the lagging (running in the 3' to 5' direction) strands (1,21,22)

Following the deposition of a single RNA primer at the DNA replication origin, the leading strand undergoes continuous, uninterrupted DNA replication. By contrast, the lagging strand undergoes discontinuous DNA replication, where multiple RNA primers are laid down and short DNA Okazaki fragments (~100-200bps) are synthesised, which are ultimately ligated together. For both the leading and lagging strand, once DNA synthesis has taken place, the RNA primers are replaced by DNA (1,22).

Finally, DNA replication enters the termination stage, where it is believed that two adjacent DNA replication bubbles meet (Fig1.1B). The remaining RNA primers are replaced with DNA, DNA polymerase is removed, and a ligase enzyme facilitates joining the newly synthesised DNA fragments together. It is important to note that the precise mechanisms by which DNA replication termination takes place is currently unclear (1,23,24).

### *Chromatin in the cell cycle and DNA replication*

To fit genomic DNA into the nucleus, it is packaged into chromatin, which also plays a pivotal role in DNA replication. Chromatin must be rearranged dynamically to allow the replication machinery access to the DNA for synthesis. The chromatin environment must also be accurately replicated on newly synthesised DNA. Chromatin determines the level of compaction of regions of the genome and is responsible for dictating transcriptional activity levels (25–27).

The basic repeating unit of chromatin is the nucleosome core unit which consists of ~147bp DNA 1.7 times wrapped around a histone octamer. Each nucleosome is separated by a variable length of linker DNA, and may be associated with a linker histone, H1 (27,28). The core histone octamer is assembled by histone chaperones, where the tetramer of histone H3 and H4 (H3-H4)<sub>2</sub> is assembled first, followed by two dimers of histones H2A and H2B (26,29).

In order to perform necessary functions, such as DNA replication and transcription, the nucleosome can be rearranged once it is formed. This rearrangement is carried out by chromatin remodelling factors or complexes with functions that include the reversible post-translational modification, such as acetylation, methylation and ubiquitination, of histone tails. In addition, multiple histone variants can be exchanged for their appropriate counterpart. For example, the histone H2A variant, H2A.Z has been implicated in the activation of human DNA replication initiation origin sites (26,27,30–32). In section 1.3, I will discuss the histone variants and histone marks (ie histone tails with post-translational modifications) that have been implicated in DNA replication origin specification and/or activation.

During DNA replication, the chromatin landscape and its rearrangement play a crucial role. The degree of chromatin compaction influences DNA replication origin activation. The

transcriptionally active and decondensed euchromatin replicates in early S-phase, while transcriptionally silent and compacted heterochromatin replicates in late S-phase (27,33).

It is believed that chromatin must be rearranged during DNA replication in order to allow the DNA replication machinery access to all genomic DNA, including DNA wrapped around nucleosomes (27,34,35). The ATP-dependent chromatin remodellers, NuRD, INO80, ISW2 and ASF1, have been shown to facilitate nucleosome remodelling at replication forks (27,34,36,37). In particular, the MRX complex, chromatin modifiers Gcn5 and Set1, and histone remodellers RSC, CHD1 and ISW1 all act to stimulate chromatin remodelling at previously stalled replication forks (38). Histone deacetylases (HDACs) 1/2 are often directed to DNA breaks and chromatin at replication forks and their removal reduces cell proliferation and reduces replication fork velocity (39–42).

HDACs and ATP-dependent chromatin remodellers may act together in DNA replication. A HDAC target, H4K16ac, acts as an inhibitor for the ISWI-family ATP-dependent chromatin remodeller, SMARCA5. SMARCA5 associates with nascent DNA and its depletion results in a similar reduction in replication fork velocity to that seen by HDAC1/2 inhibition. During early S-phase, levels of H4K16ac associated with DNA replication origins reduce, while levels of SMARCA5 increase. This suggests that deacetylation by HDAC may facilitate the activity of ATP-dependent chromatin remodellers' during replication (39,42–44). Moreover, it has also been suggested that chromatin remodellers and histone chaperones, such as ASF1 and FACT, may act together during replication fork progression (27,45–47).

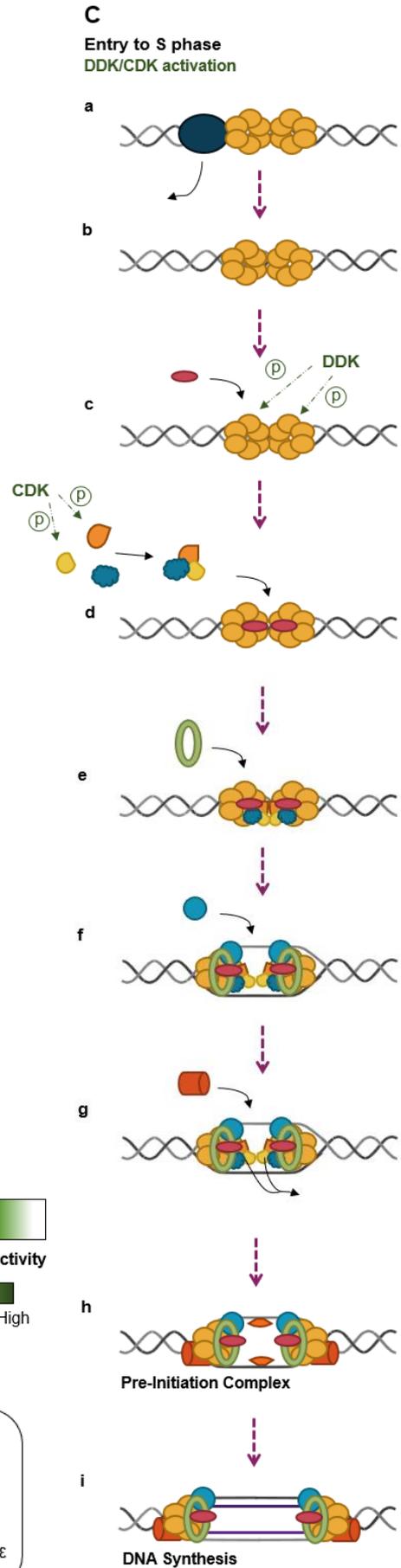
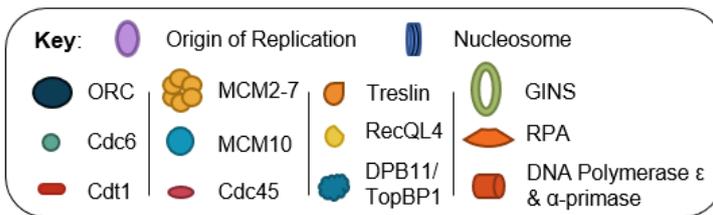
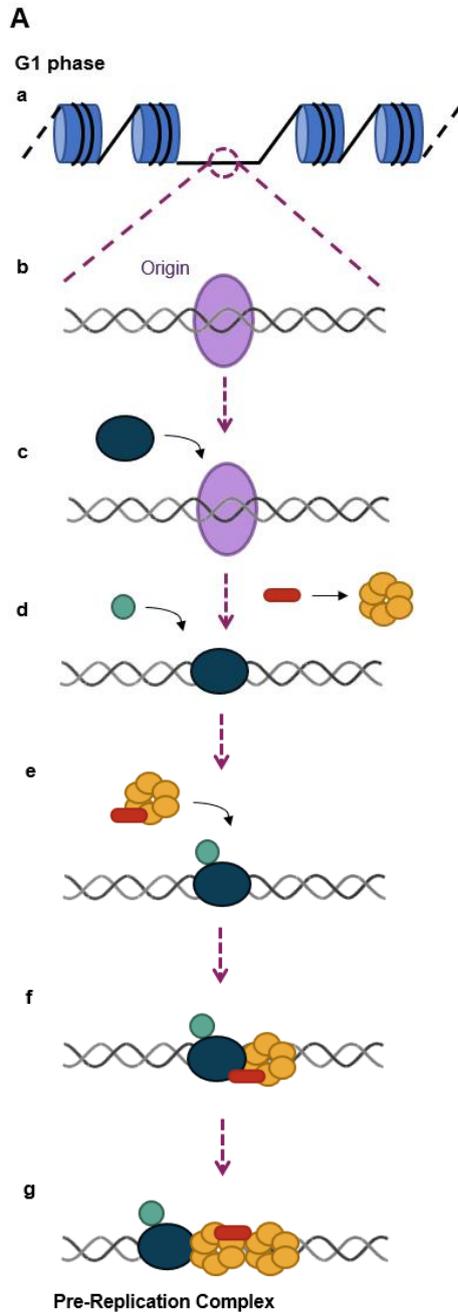
ASF1 and FACT are histone chaperones that interact with subunits of the replication fork protein, MCM2-7 complex, during DNA replication; MCM2 can act as a histone chaperone for both H3-H4 dimers and tetramers. These chaperones are responsible for disassembling the nucleosome and sequestering parental histones, during replication. After DNA replication has taken place, the nucleosomes must be rapidly reassembled on the daughter DNA strands. Additional histone chaperones, including CAF1 and NAP1 facilitate the formation of these nucleosomes. The three-subunit chaperone, CAF1 is regulated by its phosphorylation by DDK, and once phosphorylated, it associates with the PCNA-bound DNA (PCNA remains bound to the newly replicated DNA for up to 20 minutes) and deposits the H3-H4 tetramer. ASF1 donates the H3-H4 tetramer to CAF1 during nucleosome assembly. Whereas, NAP1 deposits the H2A-H2B dimer, and FACT acts as the donor of H2A-H2B, following the deposition of the H3-H4 tetramer (27,47–50). Additionally, the core DNA polymerase of the leading strand, pol  $\epsilon$  can bind H3-H4, which is mediated by two non-catalytic subunits, POLE3 and POLE4. RPA can also bind H3-H4. The DNA polymerase, pol  $\alpha$ , has been found to bind with H2A-H2B in yeast and preferentially bind to H3-H4 in mice and yeast. The

mutant defect of the histone binding region of pol  $\alpha$  has been found to impair parental H3-H4 transfer to the lagging strand (50,51).

## **1.2 Eukaryotic Chromosomal DNA Replication Initiation**

The main regulator of DNA replication is at the initiation stage. The mechanism of DNA replication initiation ensures that DNA replication origin sites are “fired” or activated only once per cell cycle and prevent re-replication (20). The DNA replication origins become associated with pre-RC components during the late G1-phase and once associated, these origins are considered ‘licensed’ for DNA replication (Fig1.2A) (15,52).

The pre-RC comprises the Origin Recognition Complex (ORC), Cell division cycle 6 (Cdc6), Cdc10-dependent transcript 1 (Cdt1) and MiniChromosome Maintenance proteins/complex 2-7 (MCM2-7) (53). The ORC, Cdc6 and Cdt1 proteins are vital for MCM2-7 complex’s association with the DNA. Once MCM2-7 has been stably associated with the DNA, ORC, Cdc6 and Cdt1 proteins are no longer required and are ultimately removed from the DNA before elongation takes place (15,53). ORC associates with the replication origins in an ATP-dependent manner (Fig1.2Aa), and then acts as a scaffold for the assembly of subsequent replication initiation factors. ORC recruits Cdc6 and Cdt1 through independent pathways to the DNA (15,53–55). Cdt1 first associates with the MCM2-7 hexamer complex (56) and both are then recruited to the ORC bound DNA (Fig1.2Ab-g). Once the first MCM2-7 hexamer is loaded, Cdc6 and then Cdt1 are released. A second Cdc6 is subsequently recruited to ORC and the second MCM2-7 associated with a second Cdt1 are recruited; Cdc6 and Cdt1 are then sequentially released. It is believed that when the second MCM2-7 is stably and completely loaded onto the DNA, ORC is simultaneously removed (Fig1.2Ca). Whether the loading of two MCM2-7 complexes required one or two ORCs present at the origin remains to be elucidated (57,58).



**Figure 1.2:** (A) The schematic summary of the formation of the pre-replication complex (preRC) during G1. (a) DNA packaged in nucleosomes (blue cylinder) and a potential origin are highlighted. The following proteins are loaded to the replication origin in the following order: (c) OCR, (d) Cdc6, and (e) Cdt1 associated with MCM2-7. (f) The loaded first MCM2-7, following which, a second MCM2-7 complex is loaded, and the pre-replication complex is formed. (g) A second MCM2-7 is subsequently loaded. (B) The activity levels of the protein kinases, DDK and CDK (green), throughout the cell cycle. Origin licensing (ie preRC formation) and origin firing are also indicated. (C) The schematic summary of pre-initiation complex formation during S phase. (a) Following preRC formation, Cdc6 and Cdt1 are removed, leaving ORC and MCM2-7 at the origin. (b) ORC is also quickly removed following the addition of the second MCM2-7. (c) Cdc45 is loaded and DDK phosphorylates 5 of 6 MCM2-7 subunits. (d) CDK phosphorylates treslin and RecQL4 that results into the interaction between treslin, RecQL4 and DPB11/TopBP1, which ultimately associates with the origin. (e) The GINS complex is recruited, at which point MCM2-7 helicase becomes activated and then MCM10 is loaded (f), which enhances the MCM2-7 helicase activity. Finally, DNA polymerase  $\epsilon$  &  $\alpha$ -primase are subsequently loaded to the pre-initiation complex (g). RPA associates with any single-stranded DNA (h). Subsequently, DNA synthesis is undertaken (i).

ORC is a 6-subunit complex and a member of the AAA+ family of ATPases. During its recruitment of MCM2-7, ORC undergoes a conformational change from a flat ring to a right-handed spiral ring. It is responsible for the selection of the replication origin sites, which may ultimately be activated for DNA replication initiation (55,57,59–62). ORC associates with a larger number of sites that are typically fired during replication initiation, indicating that there is an/are additional factor/s that regulate the firing of replication origins (63). In yeast, ORC associates with specific ori element sites, whereas, in higher eukaryotes such as humans, no such sequence specificity is present (55,64). The recognition and association of ORC to the appropriate origin sites, requires its binding to but not hydrolysis of ATP (15,54).

The Cdc6 protein is also an AAA+ family member and is another essential regulator of DNA replication initiation (65). In addition to its role in DNA replication Cdc6 is also involved in the regulation of checkpoint mechanisms that coordinate S- and M-phase (66). Due to its unstable nature, Cdc6 is rapidly degraded at the late G1 to S-phase transition and must be newly synthesised ahead of the formation of the pre-RC for each cell cycle (67–69). It binds with ORC associated DNA and may activate ORC's potential as a 'molecular switch' that then contributes to pre-RC formation. Once in place, Cdc6 facilitates the recruitment of MCM2-7 to the pre-RC (70–72). The overexpression of Cdc6 has been associated with a variety of cancers and is believed to promote DNA hyper-replication in various cell lines (66,73).

The Cdt1 protein is also recruited to ORC associated sites of DNA, however this is independent of Cdc6 (15,53,74). Unlike ORC and Cdc6, Cdt1 lacks enzyme activity, however, it does possess a cyclin binding domain (75) and associates with the MCM2-7. Cdt1 binding to MCM2-7 augments its binding- and helicase- activity (20,76), facilitating the

loading to MCM2-7 to the pre-RC through the recruitment of the Cdt1-MCM2-7 complex to ORC bound sites (56). Cdt1 has been found to be essential for cell cycle control, normal organism development and genome stability (77), and is the target of the DNA replication initiation inhibitor, geminin (78).

The MCM2-7 complex is a toroidal 6-subunit helicase (although this is a weak activity), that requires ATP binding and hydrolysis to catalyse the unwinding of the double-stranded DNA during replication; it provides the motor activity for duplex unwinding. This toroidal structure resembles that of the well-studied homo-hexameric helicases of prokaryotes (79–82), but the eukaryotic MCM2-7 complex is comprised of 6 unique but highly related subunits. Each of these subunits possess an ATPase activity although they all have their own distinct function (15,18,61,83). It is predicted that two MCM2-7 complexes associate with the pre-RC and acts at the replication fork, as the abundance of MCM2-7 at the replication fork is greater than that which would be expected if a single complex was present (15). Once assembled as part of the pre-RC, the MCM2-7 complex remains inactive until the transition to S-phase (84), where the pre-initiation complex is formed (85). In particular, the MCM2-7 complex is activated upon binding of Cell division cycle 45 (Cdc45) and GINS (86).

Upon the transition to S-phase, the activity levels of protein kinases, DDKs and CDKs, increase and DNA replication initiation is activated (Fig1.2B) (87). Ultimately the phosphorylation of multiple protein by these kinases stimulates the binding of Cdc45 and GINS to MCM2-7, forming the pre-initiation complex (Fig1.2C) (15–17). The pre-RC is formed when the CDK and DDK levels are low, and the helicase activation, pre-IC formation and subsequent replication initiation requires the phosphorylation of various proteins; it has been suggested that these differences in kinase levels acts to prevent re-replication as the kinase levels following initiation are not conducive with the formation of additional pre-RCs (15,87–89).

The Dbf4-dependent kinase (DDK) is localised to the chromatin at replication origin sites, where it associates with MCM2-7 and phosphorylates five of its six subunits. The phosphorylation of MCM4 is required for the loading of Cdc45 to MCM2-7. DNA polymerase  $\alpha$  and the pre-initiation complex protein, Cdc45 may also be targets of DDK (16,71,90). Whereas, the cyclin-dependent kinase (CDK) interacts with ORC and Cdc6 at the replication origin (91). In yeast CDK is known to phosphorylate Sld2 and Sld3, leading to their enhanced interaction, along with an additional protein, Dpb11; during DNA replication stress, this interaction between Sld3 and Dpb11 is blocked by the Rad53 kinase checkpoint (92). The human homologues of Sld2, Sld3 and Dpb11 are RecQL4, Treslin and TopBP1 respectively and are believed to play similar roles as their yeast counterparts (93). Furthermore, the association of Sld3 and Dpb11 is mirrored by that of Treslin and TopBP1, where the CDK-

mediated phosphorylation of Treslin leads to its interaction with TopBP1 (94,95). DDK and CDK activity result in the association of Treslin and Cdc45, and Treslin's association with RecQL4 and TopBP1 through the Treslin-TopBP1 interaction (88,96,97). In yeast, RecQ4 is required for stably GINS loading to MCM2-7. However, in *X. laevis*, this appears to not be the case (57,98).

The activation of initiation, instigated by DDK and CDK kinase activity, ultimately leads to the recruitment of Cdc45, GINS and MCM10 to MCM2-7, forming the Cdc45-MCM-GINS (CMG) helicase complex and eventual activation of the MCM2-7 helicase activity and recruitment of the DNA polymerases (15,16,20).

The process begins with the recruitment of Cdc45 to MCM2-7 (Fig1.2Cc). Subsequently, Treslin, RecQL4 and TopBP1 are recruited (Fig1.2Cd), followed by the GINS complex (Fig1.2Ce). Then MCM10 is loaded (Fig1.2Cf) and finally the DNA polymerases  $\epsilon$  and  $\alpha$ -primase are recruited and Treslin, RecQL4 and DPB11 are removed (Fig1.2Cg) (15,16,20,57,99–101). This leaves the components of the CMG complex and DNA polymerases which encompasses the DNA replication machinery (Fig1.2Ch). This machinery travels along the DNA at the replication fork, replicating the parental strands. Where any single-stranded DNA (ssDNA) is present, replication protein A (RPA) binds to it, preventing the formation of secondary DNA structures (Fig1.2Ci) (15,57,101,102).

Cdc45 is required for the loading of the DNA polymerases to the replication machinery and appears to be crucial in the loading of RPA onto nascent ssDNA (103–105). Cdc45 appears to be the rate limiting component of the CMG complex, in the firing of origins and unwinding of the duplex. In fact, Cdc45-overexpression results in the firing of at least twice as many DNA replication origins but a reduction in the fork elongation rate (106).

GINS is a ring-shaped, four subunit complex comprised of the Sld5, Psf1, Psf2 and Psf3 proteins (107). Together with the Cdc45 protein, GINS enhances the helicase activity of the over CMG complex (108). GINS has been found to mediate multiple interactions of many other replication factors (107,109). Furthermore, the CMG complex may be an accessory factor to the DNA polymerase  $\epsilon$  (16).

The protein minichromosome maintenance 10 (MCM10) uses an unknown mechanism to drive the initial DNA unwinding at origin sites (110), enhance the helicase activity of MCM2-7 (111) and possesses its own putative primase activity (112). MCM10 is thought to be required for the separation of the two MCM2-7 proteins prior to DNA replication elongation beginning (113). It interacts with ORC but is not dependent on it for chromatin binding (15,114). MCM10 also binds to MCM2, acting to maintain MCM2's chromatin association (15) and stabilise the MCM2-7 complex's association with Cdc45 and GINS (110).

Additionally, MCM10 has been found to be involved in Cdc45 and RPA recruitment (115). MCM10 has been suggested to function as an activator of the CMG complex during replication; it has also been implicated in the stimulation of DNA replication both *in vitro* and *in vivo* (110). Finally, MCM10 and the homotrimer, chromosome transmission fidelity 4 (Ctf4) are believed to contribute to the recruitment of DNA polymerase  $\alpha$  and Ctf4 may then act as a bridge between the CMG complex and DNA polymerase  $\alpha$  (57,116–118).

The DNA polymerases that are finally loaded onto ssDNA for bulk DNA synthesis are DNA polymerase  $\epsilon$ ,  $\delta$  and DNA polymerase  $\alpha$  primase (15,101,119). Polymerase  $\epsilon$  is loaded first and is necessary for the recruitment and loading of polymerase  $\alpha$  primase (15,101). As indicated by its name, polymerase  $\alpha$  primase possesses the primase activity which lays down the RNA primers that are essential for DNA synthesis (120,121). These RNA primers are then used by DNA polymerases  $\alpha$ ,  $\delta$  and  $\epsilon$  and DNA synthesis takes place (119). As the replication machinery moves away from the replication origin sites, DNA replication enters the elongation stage (15).

The highly conserved RPA protein is a stable heterotrimer consisting of RPA1, 2 and 3 (122). It is vital in many DNA metabolic pathways including DNA replication and repair (123). RPA binds to the backbone of the ssDNA in order to prevent DNA folding and the formation of secondary structures (124,125). Additionally, proliferating cell nuclear antigen (PCNA) is loaded by replication factor C (RFC) to the primer template laid down by the primase and acts a ring-shaped sliding clamp for DNA polymerase  $\delta$  and  $\epsilon$ , which tethers them to the DNA and ultimately enhances their processivity (57,101,118).

Together, these proteins and complexes bring about DNA replication initiation. However, they form only part of the story and the precise mechanism by which DNA replication initiation is brought about remains only partially elucidated (126). There are additional known factors that are essential for DNA replication initiation, such as Y RNAs, whose roles are not fully understood. These Y RNAs have been isolated from the cytosolic extract of human cells (human cytosol) and shown to be essential for DNA replication in both *in vitro* and *in vivo* (127). In addition, there may be further additional and, as yet, unknown factors involved in replication initiation (126). Finally, evidence has implicated the chromatin environment in the specification of DNA replication origin sites and the initiation of DNA replication (88).

### **1.3 Eukaryotic Chromosomal DNA Replication Origins**

The mechanism of DNA replication detailed above (1.2) demonstrates the role of the highly conserved protein machinery, known as *trans*-acting initiator factors. Of course there is another element at play; the sites at which these initiators act, known as DNA replication origins (aka *cis*-acting replicators). The replicon model has postulated that DNA replication at

individual sites (a replicon) consists of both *trans*-acting initiators and *cis*-acting replicators. Although many initiators, such as ORC, have been identified, the replicators have been elusive (88,128,129).

In budding yeast, potential replication origins are indicated by a shared 11bp, AT- rich consensus sequence, known as the autonomously replicating sequence (ARS). The initiator ORC protein recognises and binds with this ARS sequence to bring about DNA replication initiation (130). The specification of DNA replication origins in higher eukaryotic organisms however, is far more complex than that of budding yeast.

In metazoa, DNA replication origins are highly heterogenous, where they do not share a consistent primary DNA sequence consensus (20,88). It has been observed that not all potential origins fire in all cells of each cell cycle; in particular, the origins that initiate replication differ by up to 50% in different cell types (131–135). Additionally, for most metazoan somatic cells, only 10-20% of potential origins are fired and initiate DNA replication, thus indicating a degree of flexibility in initiation patterns, probably resulting from differing stochasticity of activation of the origin sites. Where a potential origin is not fired and remains dormant, they then undergo replication from adjacent, neighbouring replication origins (88,132,136,137).

Currently, there are multiple characteristics, ranging from DNA sequence to epigenetic factors that are associated with human DNA replication origins. However, the precise characteristics that are necessary and sufficient for the specification of origin sites remains unclear (88,129).

With the advent of next generation sequencing (NGS), it has become possible to investigate origins on a genome-wide scale. Multiple NGS methods have been developed and have identified tens of thousands of origins. The established methods that utilise NGS include: small nascent strand sequencing (SNS-Seq), replication bubble sequencing (bubble-seq), Okazaki fragment sequencing (OK-seq), initiation site sequencing (iniSeq), and Repli-seq.

SNS-seq has identified 50,000 to 250,000 discrete potential human origin sites through the sequencing of short nascent DNA strands (131,132,138). SNS-seq utilises the isolation and purification of nascent RNA-primed DNA synthesised on the leading strand, at DNA replication origin sites. These nascent strands are isolated through heat denaturation and purified from the Okazaki fragments through size fractionation. Any contaminating DNA fragments (DNA without a 5' RNA primer) are then degraded by lambda exonuclease, and the resultant purified DNA is sequenced (139,140). The potential origin sites identified by SNS-seq are often present near/at transcription start sites (TSS), G-quadruplexes (G4) and CpG islands (CGIs). It is important to note that SNS-seq is known to have a bias for GC-rich

regions of DNA as G4s are known to be resistant to lambda exonuclease digestion (131,139,141,142).

Bubble-seq sequences DNA replication bubbles and has identified > 100,000 potential origin sites (143). During the replication of genomic DNA, replication bubbles are formed when two replication forks undergo elongation. Bubble-seq utilises this early replication intermediate, by fragmenting early replicating genomic DNA (through restriction endonucleases) and embedding the fragmented DNA into polymerising agarose. Once fully polymerised, the agarose gel undergoes electrophoresis. The replication bubbles are 'trapped' in the gel, by the agarose fibres, and the linear DNA and Y-shaped replication fork fragments run out of the gel. The replication bubbles are then extracted from the gel and sequenced (144,145). As with SNS-seq, these origins were often associated with TTS of active genes (143).

OK-seq isolates and sequences Okazaki fragments and has identified 5,000 to 10,000 broad origin sites of up to 150Kb in size (146). Okazaki fragments are enriched by the inactivation of the DNA ligase I (ligates neighbouring Okazaki fragments together), isolated from asynchronous cells and sequenced. Due to the strand-specific nature of Okazaki fragments, potential origin sites can be determined from the transition in strandedness of the fragments (146,147). Discrete origin sites (within 5Kb in size) are indicated by a sharp transition, for which 66 origins were identified. OK-seq predominantly identified much broader initiation zones, with gradual transition of strandedness of 6-150Kb in size (with a mean of 30Kb). Additionally, it was observed that replication and transcription were co-oriented in the same direction. As with SNS-seq and bubble-seq, the zones and sites identified by OK-seq were frequently associated with CGIs and TSSs (129,146).

SNS-seq, bubble-seq and OK-seq have poor concordance with one another; 33 – 65% of origin sites identified through SNS- and bubble-seq overlap. Whereas, the concordance between the broad zones identified by OK-seq and large bubble-seq origin sites was greater than that of SNS- and bubble-seq. In order to reconcile the discrepancies and poor concordance between the previously established method, iniSeq was developed (148).

iniSeq combines the established human cell-free system for DNA replication initiation (explained in section 1.4) and NGS. Further details about the iniSeq methodology are specified in section 1.5. iniSeq identified < 25,000 discrete origin sites, which were often associated with TSSs and G4s. However, iniSeq is biased towards identifying early replicating origins (148) (See section 1.5).

Repli-seq is a method developed from the repli-chip method, to address the origins that fire early and late in S-phase. Repli-seq pulse-labels cultured cells with bromo-dUTP (BrdUTP) (to label newly synthesised DNA). The cells are subsequently separated into early and late

S-phase fractions through flow cytometry (depending on DNA content). The BrdU-labelled DNA is then immunoprecipitated from the early and late S-phase fractions, fragmented, sequenced and a ratio of nascent DNA from early versus late S-phase fractions is produced (149). Repli-seq can discern the temporal coordination of DNA replication (150).

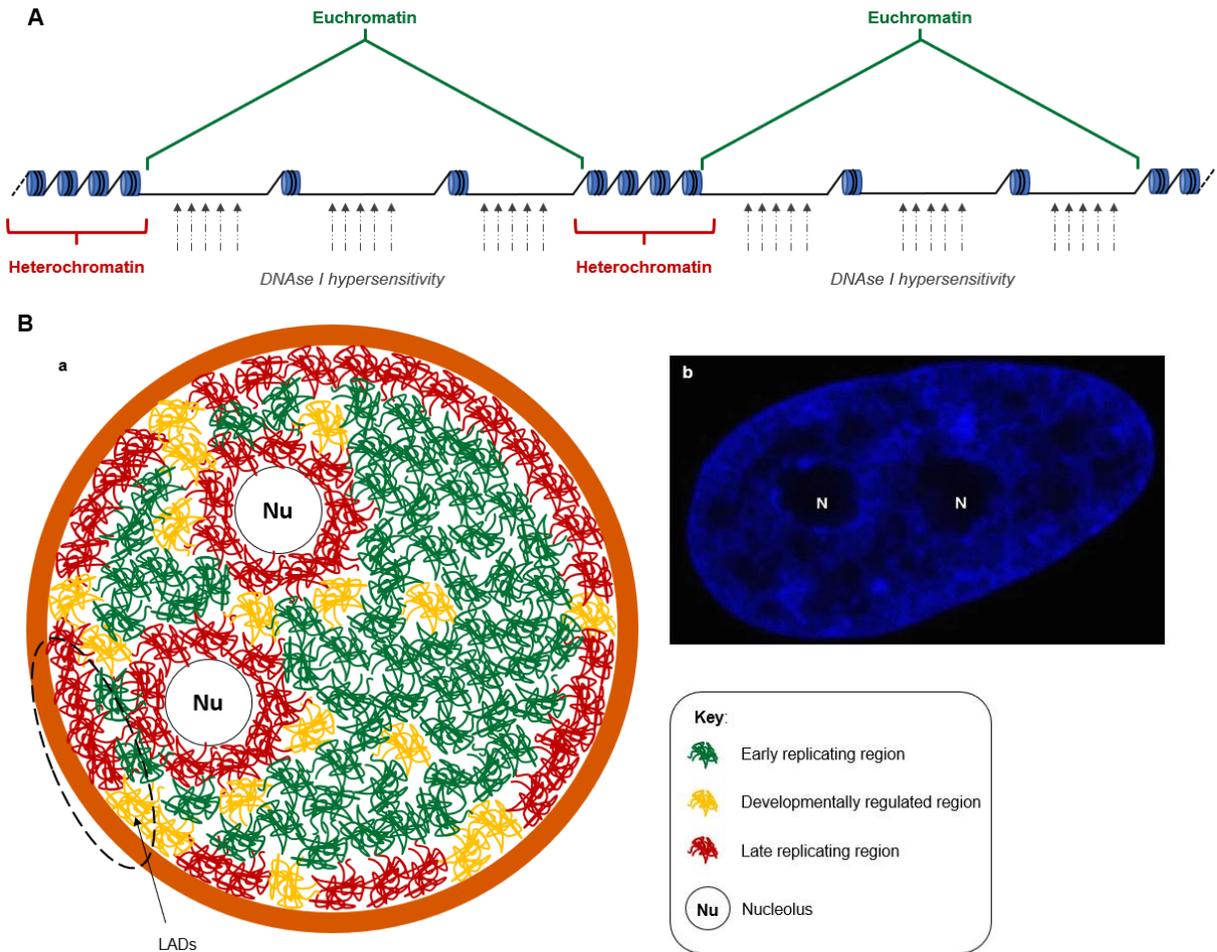
Taken together with other sequencing methods such as Chromatin Immunoprecipitation sequencing (ChIP-seq) and Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq), an array of characteristics have been identified that are associated with the regulation, specification and activation of human DNA replication origins. These characteristics range from low resolution global nuclear architecture, through to high resolution primary DNA sequence and are described below.

### *1.3.1 Global nuclear architecture*

Genomic DNA is spatio-temporally compartmentalised into the nucleus, which possesses a global nuclear structure/architecture (151,152). Replication origins/domains have been found to associate with particular nuclear structures (153,154).

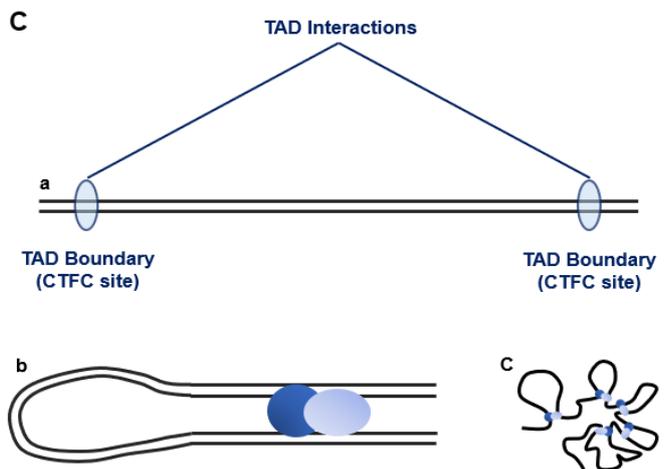
The chromatin is separated into two categories: euchromatin and heterochromatin. Euchromatin is decondensed, open and accessible. It possesses a low density of nucleosomes and is transcriptionally active. Heterochromatin is highly condensed, nucleosome abundant and transcriptionally silent. Euchromatin is present in the nuclear interior and heterochromatin is present around the nuclear and nucleolar periphery (Fig.3B) (1,155–159).

Early replicating domains are present in euchromatin and exhibit a high density of efficient origins (160). To support this, replication origins have been found to be present at DNase hypersensitive regions of DNA (132,161). This hypersensitivity has been long associated with open chromatin where the DNA is more accessible to the enzyme (Fig1.3A) (162). Additionally, immunofluorescence has identified replication foci, which are 0.5-1Mb regions of replicating DNA with multiple synchronously firing replication origins. Early replicating foci are often situated at the nucleus interior, which is consistent with the position of euchromatin (154,163).



**Figure 1.3:** (A) a schematic of euchromatin (green) and heterochromatin (red). Heterochromatin is condensed and nucleosome (blue cylinder) abundant. Euchromatin is decondensed, comprises few nucleosomes and shows DNase I hypersensitivity (grey). (B) (a) A schematic of the distribution of early and late replicating regions and developmentally replicated regions within the nucleus. (b) DAPI stain (DNA) of the nucleus of HeLa cells in interphase (133).

Euchromatin (dim) is found towards the interior of the nucleus, whereas heterochromatin (bright) is found at the nuclear and nucleolar periphery. 'N' indicated the nucleolus. Late replicating domains are located around the periphery of the nucleus and around the nucleolus (a), which is consistent with heterochromatin (b). Early replicating regions are located in the nucleus interior (a), which is consistent with euchromatin (b). Developmentally regulated regions have less distinct compartmentalisation. Lamina-associated domains (LADs) are also indicated. (C) Schematic diagram of topological associated domains (TADs). (a) TADs boundaries are located long distances from one another and are defined by CTCF insulator element sites. (b) These CTCF sites come together through long-distance chromatin interactions and form a loop. (c) Multiple loops come together to form a larger overall structure.



Conversely, late replicating domains have far fewer origins and are associated with heterochromatin. The late replicating foci are tightly clustered and are found at the nuclear periphery, nucleolus and other heterochromatin regions (Fig1.3B) (154,163).

Further to eu- and hetero-chromatin, the genomic chromatin is subcategorised into lamina-associated domains (LADs) and topologically associated domains (TADs) (164,165). LADs are present around the periphery of the nucleus, in close contact with the nuclear lamina. Late replicating origins are also associated with these transcriptionally repressed LADs (Fig1.3B) (165).

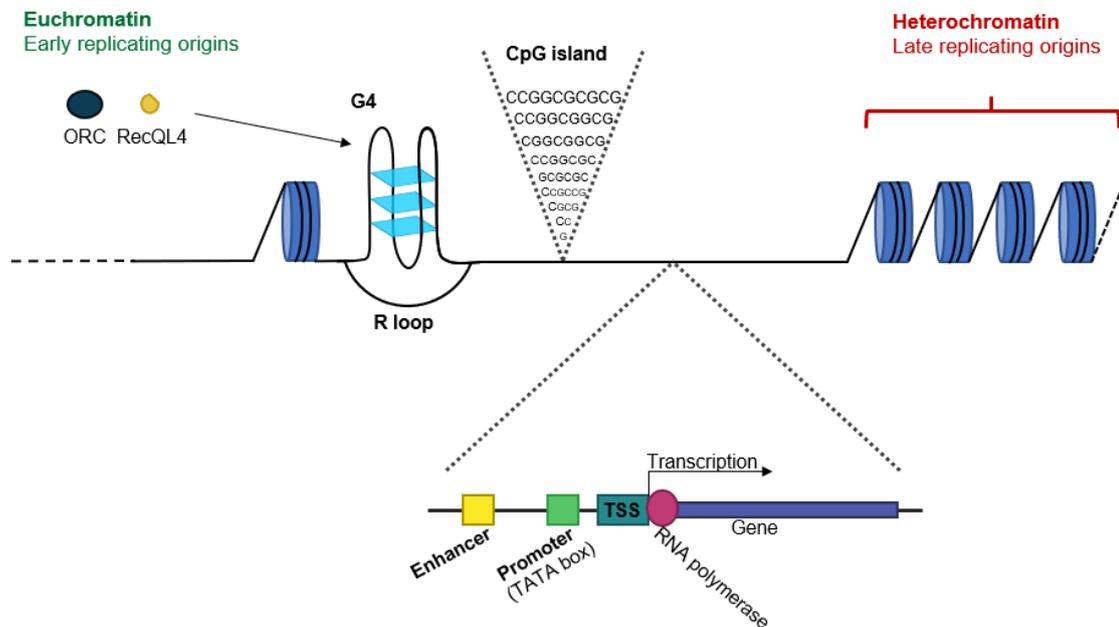
TADs are regions of DNA that are demarcated by CCCTC-binding factor (CTCF) insulator elements, which come together through long distance chromatin interactions, to form chromatin loops; these can be identified through Hi-C analysis (Fig1.3D) (164). Genome-wide analysis has shown that these TADs delineate replication timing domain boundaries (166). In fact, the deletion of the CTCF sites in Hox clusters led to the expansion of the active euchromatin into the neighbouring condensed heterochromatin. This expansion resulted in the transcriptional activation of the former heterochromatin (167). It is, therefore, conceivable that this expansion could also apply to DNA replication origins and their initiation, although this remains to be tested (129).

### *1.3.2 Primary sequence*

Despite the lack of consistent primary sequence features at metazoan replication origins, they often share some sequence similarities (88,129). Upon genome-wide analysis of the human genome, the following features have been frequently observed at/around origin sites: the GC-rich repeat elements, CGIs and G4s (and R loops); TSSs (often of actively transcribing genes), promoters and enhancers (Fig1.4) (141,168,169).

Unlike budding yeast, where origins are present at AT-rich sites, human replication origins show a preference for GC-rich DNA regions (170). CGIs are approximately 1kb long, methylation-free GC rich regions (containing >50% GC) that often appear at promoter regions (171). Small nascent strand accumulation was observed at CGIs; these have also been detected at, and not flanking, identified replication origins. Additionally, DNA regions that contain many CGIs replicate in early S-phase (172). Nascent strand accumulation has established that CGIs were present at or near greater than half the identified human origins (173). Further analysis has suggested that origins associated with CGIs are regularly more efficient than origins not associated with CGIs (141). However, CGIs are found at 60-70% of vertebrate gene promoters, which makes it difficult to distinguish the role of CGIs in DNA replication from those in RNA transcription (174).

When a DNA sequence contains four sets of three or more Guanines separated by short linker regions, a compact four-stranded helical structure is formed, known as a G-quadruplex (G4). The highly thermodynamically stable, G4s are established when the guanine motifs form a quartet structure and two or more quartets are stacked on top of one another around a monovalent cation (170,175,176).



**Figure 1.4:** Primary sequence features associated with DNA replication origins. Early replicating origins are associated with G-quadruplexes (G4s) and R loops, CGIs and transcription start sites (TSS), enhancers and promoters (some of which contain a TATA box). G4s are recognised by ORC and RecQL4 replication proteins.

Origins have been found approximately 25-300bps downstream of potential G4s (177). Based on current mapping experiments of the human genome, potential G4s have been identified at around 80% of replication origins (178). However, there are currently >370,000 predicted G4s in the human genome, which is a significantly greater number than replication origins, and would seem to imply that G4s do not constitute the sole characteristic defining origins (129).

The importance of G4s in replication origin specification has been demonstrated through the disruption of the prominent G4 present at the origin located in the chicken  $\beta^A$  globin promoter. This disruption leads to the reduction in origin efficiency and delayed origin firing. Similarly, the insertion of the same origin at a genomic location devoid of highly efficient origins (replicates during mid S-phase), of chicken DT40 cells, found enrichment of small nascent strands, but not advancement in replication timing (176,179). Additionally, the orientation of the G4 motif is believed to influence the position of the origins (168,176).

G4s have also been identified as a recognised binding site for the replication proteins, ORC and RecQL4, suggesting a role for G4s in pre-RC formation (180). However, *in vitro* replication systems in *Xenopus laevis* demonstrated that G4s are not involved in pre-RC formation, but are involved in origin firing (177). Taken together, it appears that G4s play an important, yet complex role in DNA replication origin specification and firing. Potential roles include; responsibility for nucleosome positioning, interactions with pre-RC proteins and/or facilitating the formation of R-loops (132,181,182).

R-loops form opposite a G4; they result from the hybridisation of the ssDNA with a complementary RNA strand. Multiple R-loops in close proximity can compromise genomic stability (175,183). Unsurprisingly, R-loops have been found to colocalise with ORC binding. Although the precise mechanism by which R-loops are formed remains unclear, they could facilitate the opening of DNA and exposure of ssDNA. This exposure may provide greater accessibility for the replication machinery (175).

TSSs are the sites of the genome where RNA transcription begins and have been long associated with DNA replication origins (141). *iniSeq* determined that a little over 40% of identified replication origins overlap with TSSs (148). In fact, origins have often been known to locate with moderately active genes (133,141) and DNA replication has been found to preferentially fire at TSSs of genes with high RNA polymerase II occupancy levels (184). However, the association of origins with active genes appears to lessen in highly transcribed DNA regions (133).

It remains the case that early firing replication origins are associated with regions of DNA that are transcriptionally active. Late firing origins are often associated with transcriptional silence (185). Current conjecture is that transcription and replication frequently initiate at the same sites, in order to regulate and coordinate both processes and ultimately prevent collisions and disruptions of the replication and transcription machineries (88,133,141,186).

Finally, promoters and enhancers are gene regulatory features. Promoters and enhancers are located upstream of the TSSs (187,188). Promoters often contain a TATA box sequence. Promoters have been consistently found to associate with replication origins and have been known to localise with CGIs. Interestingly, origin density has been found to strongly correlate with promoter density in mouse embryonic stem cells (141,189). Enhancers have also been identified near replication origins (169).

While these features are consistently found near or within DNA replication origins, their association remains, by and large, correlational. Moreover, it is difficult to distinguish which features are required for DNA replication, RNA transcription or both, as both processes occur spatially together (129). Evidence suggests that there is temporal separation of transcription

and DNA replication. However, this has been known to break down and result in transcription-replication conflicts (129,190,191).

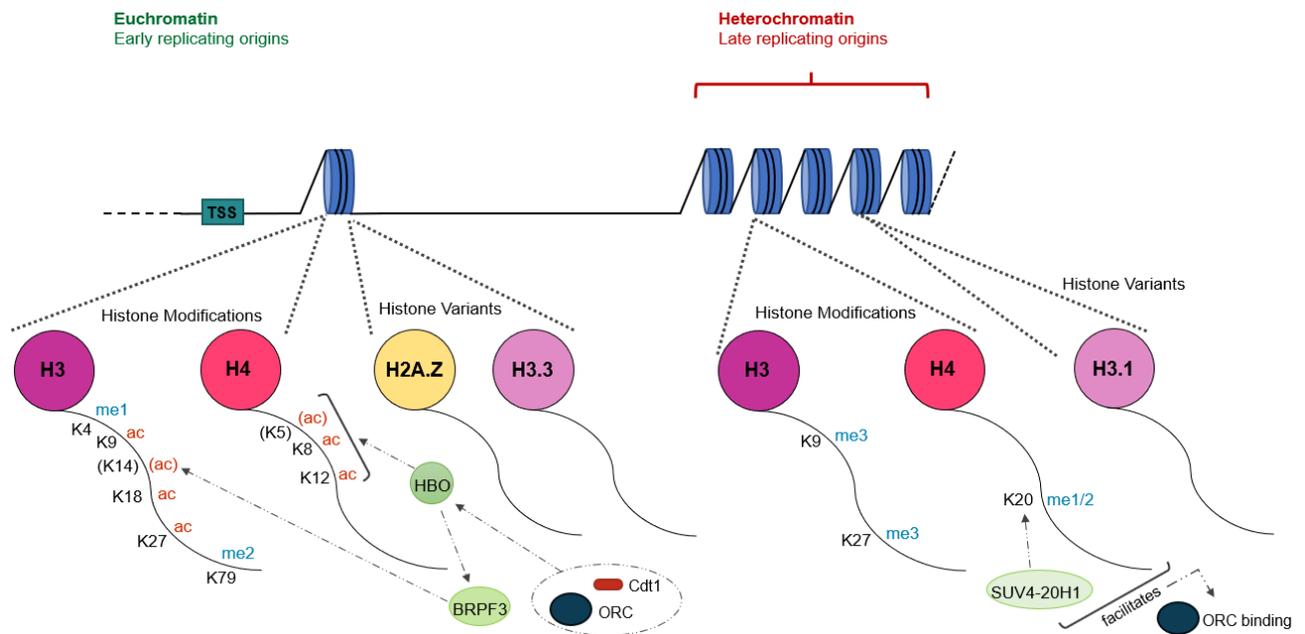
### *1.3.3 Epigenetics*

The lack of a consistent primary consensus sequence at human replication origins tends to indicate that there are additional features that play a substantial role in defining DNA replication origins. It is presently believed that epigenetics control the origin specification and activation, due to the observed flexibility and plasticity of human origins (88,129).

An extensive amount of research has been and is currently being undertaken to establish which epigenetic factors are responsible for replication origin specification. At present a variety of histone post-translational modifications and histone variants have been identified at actively replicating origin sites.

Many histone marks have been previously associated with open chromatin and active transcription (192). There also appears to be a difference in the histone marks associated with origins that fire in early and late S-phase, which may be unsurprising as early replication origins are often found in euchromatin whereas late firing origins tend to be found in condensed heterochromatin (88,129,193). Generally, early replicating origins are associated with histone acetylation and late replicating origins are associated with hypoacetylation (194–197). It is important to note that there is conjecture within the field as to the nature of the cause of origin firing timing. One model suggests that origins are subjected to defined firing timing patterns, whereas the opposing model favours stochastic firing with origins that possess varying firing efficiencies. There is potential to reconcile the two models whereby the origin firing efficiencies bring about the origin firing timing patterns. Therefore, the terms early and late firing origins can be interchangeable with more or less efficiently firing origins respectively (198,199).

The histone marks that have been found to colocalise with early replicating origins/regions include H3K4me1/2/3, H3K9ac, H3K18ac, H3K27ac, H3K79me2, H4K8ac and H4K12ac (134,138,200–203). H4K20me1/2 also associate with replication origins (204). Whereas, late replicating origins have been found to associate with H4K20me3, and methylated H3K9 and H3K27 (Fig1.5) (88,205,206).



**Figure 1.5:** Epigenetic features associated with replication origins. Early replicating origins are associated with the following histone marks; H3K4me1, H3K9ac, H3K18ac, H3K27ac, H3K79me2, H4K8ac and H4K12ac. The same origins are associated with the H2A.Z and H3.3 histone variants. The protein HBO is highlighted as a key histone acetyltransferase for the H4 histone modifications and interacts with ORC and Cdt1. HBO also targets BRPF3 (which targets H3K14). BRPF3 depletion leads to reduced origin density and Cdc45 recruitments. Late replicating origins are associated with H3K9me3, H4K27me3 and H4K20me1/2 histone marks and with the histone variant H3.1. The histone methyltransferase SUV4-20H1 methylates H4K20; both SUV4-20H1 and H4K20me2 may be required for stabilising ORC binding to replication origins rather than defining replication origins.

As previously stated, acetylation of histones occurs frequently at replication origin sites (196). The importance of acetylation can be demonstrated by the analysis of the well-established origin at the human  $\beta$ -globin gene. The  $\beta$ -globin origin in erythroid cells is shown to initiate during early S-phase and correlates with histone acetylation. Conversely, the  $\beta$ -globin origin in non-erythroid cells initiates during late S-phase and possesses reduced levels of acetylation when compared to that of erythroid cells (135,197).

In addition to their correlational observations, Goren *et al* (197) investigated the role of acetylation through the tethering of histone acetylases and histone deacetylases to the  $\beta$ -globin origin in lymphocytes and erythroid cells respectively. They found that the tethering of histone acetylases in lymphocytes was sufficient to advance replication timing of this origin; origin firing shifted from late to early S-phase. Whereas the tethering of histone deacetylases in erythroid cells was sufficient to repress origin firing at this location and delay replication timing; origin firing shifted from early to late S-phase. Thus, establishing the influential role of histone acetylation in origin selection and firing (197).

Furthermore, the histone acetylase, HBO1 is responsible for acetylating histones at H4K5, H4K8 and H4K12 (Fig1.5) (207). H4K5ac has been associated with replication origins in the

plant *A. thaliana* but there is yet to be any evidence that it is associated with human replication origins (208). H4K8ac and H4K12ac have both been identified to be crucial in the decompaction/loosening of chromatin during DNA replication (209). HBO1 is known to interact with ORC and Cdt1 and it is believed that HBO1 is recruited to origins via its interaction with Cdt1 (Fig1.5), which ultimately leads to increased acetylation at these sites (202,203,210,211).

Additionally, HBO1 also interacts with the chromatin regulator, BRPF3, in order to preferentially acetylate H3K14 near TSSs (Fig1.5). Interestingly, depletion of BRPF3 leads to the induction of replicative stress and, more relevantly, the reduction in origin density and Cdc45 recruitment to chromatin (212,213).

H3K4me1 is frequently located at enhancer regions and has been implicated in BAF complex (ATP-dependent chromatin remodeller from the SWI/SNF family) binding and chromatin regulator recruitment (200). Interestingly, the BAF complex regulates transcription through altering chromatin structure and polycomb repressive complex 2 (PRC2) in the genome (214). PRC2 is the histone methyltransferase that targets another known origin associated mark, H3K27 for trimethylation (206,215). H3K4me3 has also been found near early replicating origins in euchromatin (138). H3K18ac are frequently found at TSSs and its strong enrichment has been identified as a marker for multiple cancers (216,217). H3K79me2 is found immediately downstream of TSSs (218), is strongly correlated with gene activity, and is believed to help limit DNA replication to once per cell cycle (201). Again, these marks highlight the relationship between DNA replication and transcription.

Dimethylated H4K20 has also been implicated in replication origin specification and is enriched at replication origins. H4K20me2 interacts with the BAH domain of the ORC protein. The importance of ORC's BAH domain has been established through the disruption of this domain, which has resulted in S-phase delay and reduction in origin occupancy by ORC. Similarly, abolishing ORC recognition of H4K20me2 reduces ORC occupancy of origins (204).

The H4K20 methyltransferases, PR-Set7 and SUV4-20H1/2 are required for proper cell cycle progression and origin licensing (219–221). In particular, overexpression of PR-Set7 results in re-replication and SUV4-20H1 (which catalyses H4K20me2) facilitates ORC loading onto chromatin (Fig1.5) (204,219).

Despite the enrichment of H4K20me2 at replication origins and its interaction with ORC, H4K20me2 is unlikely to be highly specific for replication origins, as this histone mark is one of the most abundant post-translational histone marks, with >80% of the total H4 population possessing this mark (204,222). Further conjecture asserts that H4K20me2 may be required

for ORC stability, rather than origin definition (223,224). Additionally, H4K20me3 is present at late firing origins to ensure the replication of heterochromatin (205).

H3K9 and H3K27 histone marks can be associated with both early and late firing replication origins; the type of post-translational modification differs (88,134). The known euchromatin marker, H3K9ac is associated with early firing replicating origins and has been implicated in the regulation between DNA replication and transcription at the Dbf4 promoter locus (138,225). Whereas, the heterochromatin-associated, compaction marker H3K9me3, is often depleted in regions of early replicating origins and enriched in regions containing late firing origins (132,138,206,226).

H3K27ac is associated with active enhancers and thus transcriptional activity, and is enriched at early replicating origins (134,227,228). Conversely, the repressive chromatin mark, H3K27me3 is associated with transcriptional silencing and is often found at late replicating origins (206). By contrast, approximately 40% of origins that fire in early to mid S-phase have been found to associate with H3K27me3 (138). In fact, the repressive H3K27me3 has been found to strongly colocalise with the activating H4K3me3 (132). There is now conjecture that these marks together represent a bivalent chromatin state, which is commonly found at developmentally regulated genes. This demonstrates the highly dynamic role that chromatin modifications play in DNA replication origin specification and activation (134,229,230).

Finally, histone variants that can be exchanged in the nucleosome and histone variants, H3.1, H3.3 and H2A.Z (Fig1.5), have been associated with replication origins (27,138,231). H2A.Z is highly evolutionarily conserved (232) and has been found to facilitate the licensing and activation of early replicating origins. Interestingly, nucleosomes that contain the H2A.Z variant also possess enriched H4K20me2 and are associated with ORC binding and increased nascent strand accumulation. The current hypothesis suggests that, as H2A.Z binds with the SUV4-20H1 methyltransferase, it facilitates H4K20me2 deposition, which in turn interacts with ORC. Additionally, origins that are associated with H2A.Z show a greater efficiency and fire earlier than other identified origins (32).

H3.1 and H3.3 are associated with late and early replicating origins respectively. Across the course of S-phase the H3.3 variant decreases in density, whereas the H3.1 variant increases in density (231). Moreover, H3.3 has also been associated with enhancers and facilitates RNA transcription through the maintenance of decondensed euchromatin, which again highlights the connection between transcription and DNA replication (233,234).

In reality, DNA replication origin specification and activation are likely to be a combination of multiple epigenetic and DNA sequence features. Moreover, as with the primary sequence

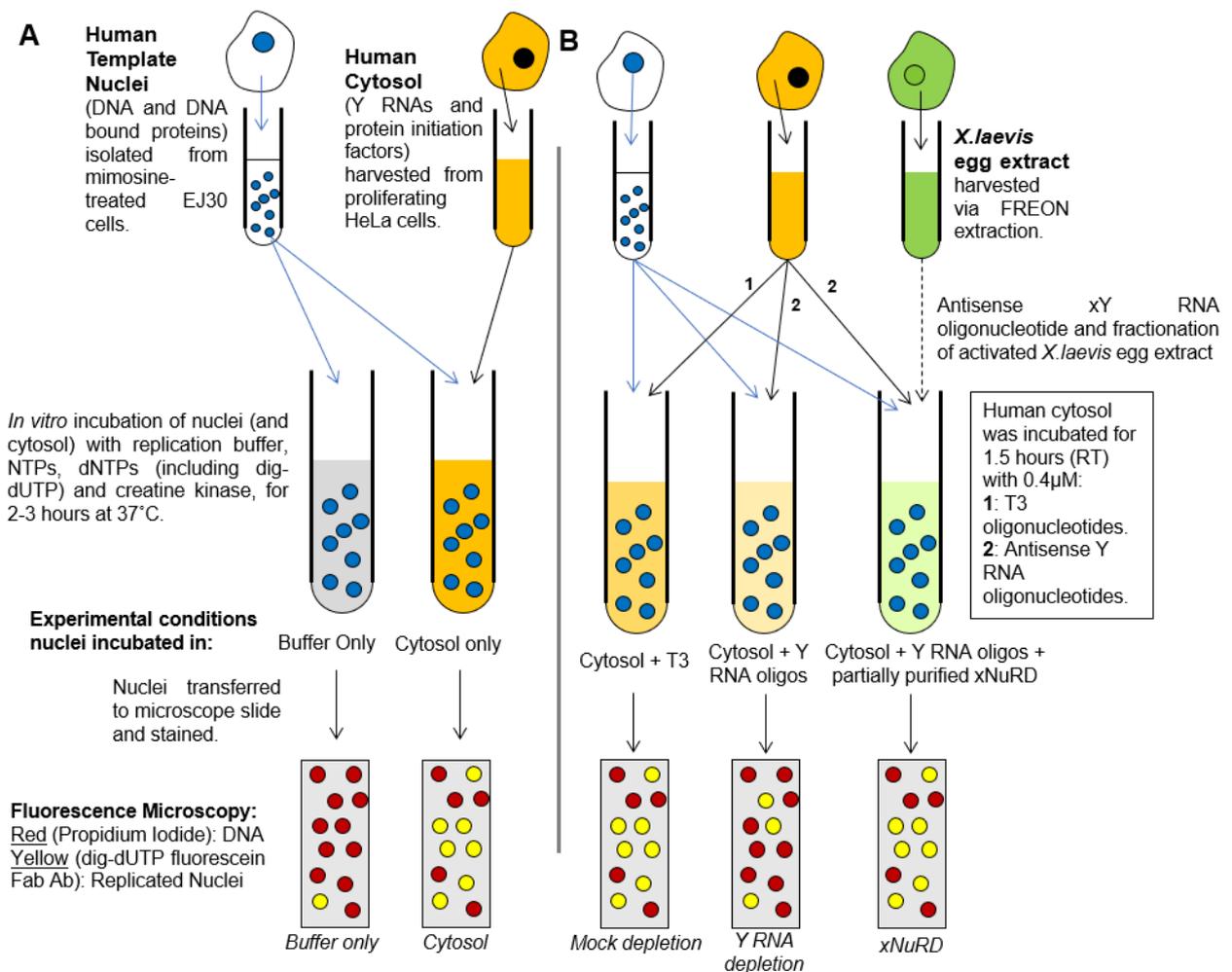
characteristics that are associated with replication origins, the vast majority of the epigenetic characteristics associated with origins are largely correlational (129). Further work remains to be conducted to determine causal links between origins and these characteristics, elucidate a mechanism by which these epigenetic and primary sequence features determine origins, and distinguish the characteristics that define DNA replication, transcription or both.

#### **1.4 Assessment of DNA Replication on the Human Cell-Free System**

The main experimental tool that underpins the foundation of the work conducted in this PhD is the *in vitro* human cell-free system which assesses human DNA replication. This experimental technique is used to determine the percentages of actively replicating template nuclei following a DNA replication reaction that allows for DNA replication initiation and partial elongation. The conditions under which the replication reactions are conducted can vary in order to determine the effect various potential DNA replication factors may have on DNA replication activity (Fig1.6A) ((235,236) and references therein).

How the human cell-free system can be used to assess DNA replication activity and be manipulated to investigate potential DNA replication factors is described below:

The human cell-free system uses template nuclei, isolated from human EJ30 cells synchronised in late G1-phase (through L-mimosine treatment), and consequently are licensed for but are not actively undergoing DNA replication (235).



**Figure 1.6:** The schematic summary of the human cell-free system. (A) shows the standard negative and positive controls, whereas (B) shows the mock (control) and Y RNA depletions and xNuRD addition. (A) In the negative control (buffer only) template nuclei that were synchronised to the late G1 phase, were isolated from mimosine-treated EJ30 cells. These nuclei were incubated with a mixture of replication buffer, NTPs, dNTP (including digoxigenin-dUTP (dig-dUTP)), creatine kinase. For the positive control (cytosol only), the template nuclei were incubated with the same constituents as the buffer only control and human cytosolic extract isolated from asynchronous, proliferating HeLa cells. (B) In the mock depletion control (cytosol + T3), the same components as described in the negative buffer only control were incubated with T3 (oligonucleotide for bacteriophage T3 DNA) treated the human cytosol. In the Y RNA depletion (cytosol + Y RNA oligos), the same components as described in the negative control were incubated with antisense Y RNA oligonucleotide treated cytosol. For the xNuRD addition, the same components as described in the Y RNA depletion condition with partially purified xNuRD. The xNuRD was purified from activated *X.laevis* egg extract, following the removal of xY RNAs (treatment with antisense xY RNA oligonucleotides). The nuclei of each condition were then transferred to individual microscope slide coverslips and stained with propidium iodide and anti-digoxigenin dUTP antibody. These are imaged through fluorescence microscopy and the nuclei are counted to provide percentages of replicating nuclei.

These template nuclei are incubated (37°C) with human cytosolic extract, a selection of NTPs (ATP, GTP, CTP and UTP), dNTPs (dATP, dGTP, dCTP and dTTP/digoxigenin-dUTP) and an ATP-regeneration system (creatine phosphate and phospho-creatine kinase), in the presence of a physiological buffer. The cytosolic extract is isolated from asynchronously proliferating HeLa cells and contain the essential DNA replication factors that are required to

facilitate the initiation of DNA replication in the licensed template nuclei (235). These template nuclei overcome the effect of the mimosine treatment within the first 10 mins and enter the DNA replication initiation stage, which is followed by DNA replication elongation (T Krude, Personal communication; (235)). The progression of the DNA replication fork is approximately 304bps/minute +/- 162bps/minute (237). Across the course of a 3-hour DNA replication incubation, >1% of the total genomic DNA within the system will undergo DNA replication (T Krude, Personal communication; (235,238)).

Upon the initiation of DNA replication, the dNTPs, including dig-dUTP, which acts as a label, assemble to form the newly synthesised DNA. Following the replication incubation, nuclei are stained with propidium iodide (for DNA) and fluorescein-labelled Fab anti-digoxigenin antibody (for incorporated dig-dUTP detection and therefore newly synthesised DNA) (235).

The nuclei are imaged using fluorescent confocal microscopy and the subsequent images are used to determine the number of replicating nuclei; red indicates nuclei and the green/yellow indicates actively replicating nuclei. The total number of nuclei and the number of actively replicating nuclei are counted and used to calculate the percentages of actively replicating nuclei under any given experimental conditions. A negative 'buffer only' (in the absence of cytosol) control is carried out alongside these experiments to establish the percentage of contaminating S-phase nuclei (235).

The major advantage of this method is the ability to manipulate the system through the addition or removal of various factors, in order to assess their potential as a DNA replication initiation factor. While the factors can be added to the *in vitro* replication reaction directly, often the removal of various factors is achieved by depleting them from the human cytosol used in the replication reactions. This method can also be modified to assess the effect of potential DNA replication factors on DNA replication elongation through the replacement of the late G1-phase template nuclei with S-phase template nuclei (235).

This experimental tool has been crucial in the elucidation of novel essential DNA replication factors. The two key examples of this use of the cell-free system is the identification of non-coding Y RNAs (127) and the chromatin remodelling and histone deacetylase complex, xNuRD as independent DNA replication initiation factors (see section 1.6) (239).

The human cytosol used in this system contains a variety of DNA replication factors, including Y RNAs (Fig 1.6B). These Y RNAs can be depleted from the cytosol using anti-sense DNA oligonucleotides. The resultant Y RNA-depleted cytosol can then be added to the cell-free system. Alongside the Y RNA depletion, an oligonucleotide complementary to the bacteriophage T3 DNA (non-human specific) is used to mock deplete the human cytosol and acts as a positive mock depletion control in the cell-free system. This accounts for the

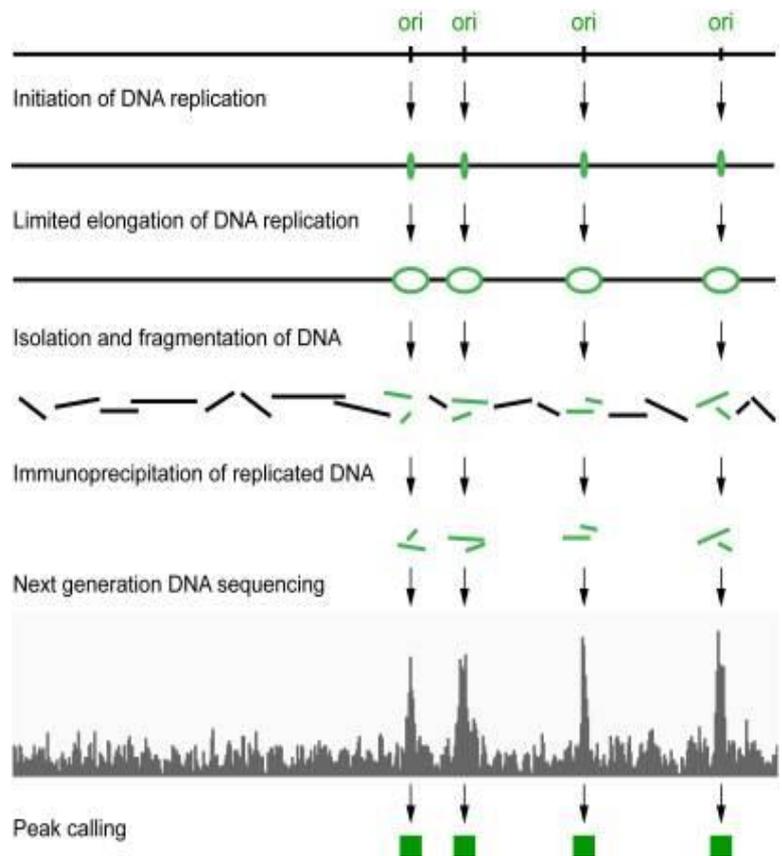
addition of a DNA oligonucleotide into the human cell-free system. When the hY RNAs are depleted/removed, there is a significant reduction in the levels of DNA replication activity, but no significant reduction in DNA replication activity levels is seen in the mock depletion control (127).

Once the hY RNAs are removed from the cell-free system (through depletion), additional factors can be introduced, to assess their effect on DNA replication. The xNuRD complex is one such factor that has been analysed in the cell-free system following removal of hY RNAs by oligonucleotide depletion. Extract from activated *X. laevis* eggs underwent xY RNA depletion, fractionation/purification and was then tested on the cell-free system. xNuRD was subsequently identified as a factor that was able to fully restore DNA replication activity in the absence of hY RNAs (239).

This powerful tool can facilitate the elucidation of the mechanism by which hY RNAs and xNuRD bring about DNA replication, including the identification of further DNA replication factor that interact with hY RNAs and assessment of their effect on human DNA replication initiation origins.

### 1.5 Initiation-Site Sequencing (iniSeq)

Since the advent of affordable next generation sequencing (NGS), the scope for high resolution analysis of DNA replication origin sites has increased substantially. In collaboration with the J.Smith group at the Francis Crick institute, the Krude group has developed the novel iniSeq approach to identify origins, based on the human cell-free system and in conjunction with NGS (Fig1.7) (148).



**Figure 1.7:** The schematic summary of iniSeq 1.0 method. A digoxigenin tagged dUTP (dig-dUTP) was incorporated into newly synthesised DNA (green) during a short 15-minute *in vitro* DNA replication reaction. The resultant genomic DNA was subsequently isolated and fragmented. The newly replicated, fragmented DNA was separated from total DNA through immunoprecipitation for anti-dig. This replicated DNA then sequenced (Illumina sequencing) and aligned to the human genome (hg19 reference genome). Finally, origin sites were called through SICER analysis. (Schematic from 126)

Digoxigenin-dUTP (dig-dUTP) is incorporated into newly synthesised DNA during an *in vitro* DNA replication initiation reaction using synchronised late G1-phase nuclei (as occurs in the human cell-free system). Following the end of the replication reaction the excess reaction constituents are washed away and the protein and RNA are degraded. The total DNA is then fractionated (100-1000bp) through sonication and the digoxigenin-labelled DNA is isolated through immunoprecipitation with an anti-digoxigenin antibody. The total DNA and the purified newly replicated (digoxigenin-labelled) DNA are then sequenced and bioinformatically analysed (148).

Langley *et al* (148) demonstrated that iniSeq was a viable and successful high-resolution (sub-kilobase resolution) method of replication origin identification. They identified over 25,000 site-specific DNA replication origins and exhibited a high concordance with biological repeats. Many of the > 25,000 origins identified by iniSeq were found to localise with TSSs and/or G-quadruplex (G4s) structures. It was proposed that the presence of the replication origins at TSSs may be to prevent the potential head-on collisions and collapse of the transcription and replication machinery (148).

The iniSeq origins showed a good concordance with replication origins identified by the established SNS-seq, OK-seq and bubble-seq NGS methods, with the highest concordance shown between iniSeq and SNS-seq. IniSeq provides the benefit of being able to overcome shortcomings in these other methods (148), which include: the SNS-seq bias for GC-rich sequences (as a result of the inefficient digestion of these regions by lambda exonuclease) (146), and the bubble-seq bias against small replicated fragments and asymmetrically located origins on small fragments (143). One disadvantage of the iniSeq method is that it possesses a bias for early replicating origins, unlike bubble- and OK-seq which also detect late firing origins (139,143,148).

Finally, the major advantage of this iniSeq method is that as it is based on the human cell-free system, the replication reactions can be manipulated to assess the effect of various factors on DNA replication initiation. Initiation factors under investigation can be removed or added and DNA structures, such as G4s, can be stabilised or disrupted in the replication reactions. These reactions can then undergo the full iniSeq analysis and provide functional correlational analysis about the role of these potential replication factors on origin specification and activation, rather than just descriptive analysis about replication origins alone (148).

## 1.6 Y RNAs and their Role in Chromosomal DNA Replication

Historically, it was believed that proteins were the sole implementers of genetic information in the cell (240). However, the discovery of the non-coding RNAs (ncRNAs), rRNA and tRNA in the 1950s and mRNAs in the early 1960s changed this assumption, and it became the common belief that mRNAs were the key regulators of cellular control (241,242).

We now understand that 80-90% of the human genome is transcribed and exhibits functional roles (243), but only 1-3% (depending on genome annotation versions) is protein coding (244). In addition, Genome-wide Association Study (GWAS) studies demonstrated that nearly 90% of all disease phenotype-associated SNPs reside within non-coding regions of the human genome (245,246).

Since these first observations of rRNAs, tRNAs, and mRNAs, which are vital for protein production, the field has expanded to include an array of new classes and sub-classes of ncRNAs that play significant roles in cellular functions (242). These ncRNAs are often divided based on RNA size; small ncRNAs (< 200 nucleotides), such as small nuclear RNAs (intron excision), small nucleolar RNAs (modifications of rRNAs) and miRNAs (key player in gene expression regulation), and long ncRNAs (>200 nucleotides) (247,248).

The discovery and analysis of a class of small ncRNAs, known as Y RNAs, is of particular note for this thesis. Y RNAs were first identified as targets for autoimmune sera of patients with the autoimmune disorder, lupus erythematosus, and in patients with Sjögren's syndrome, in the early 1980s (249,250).

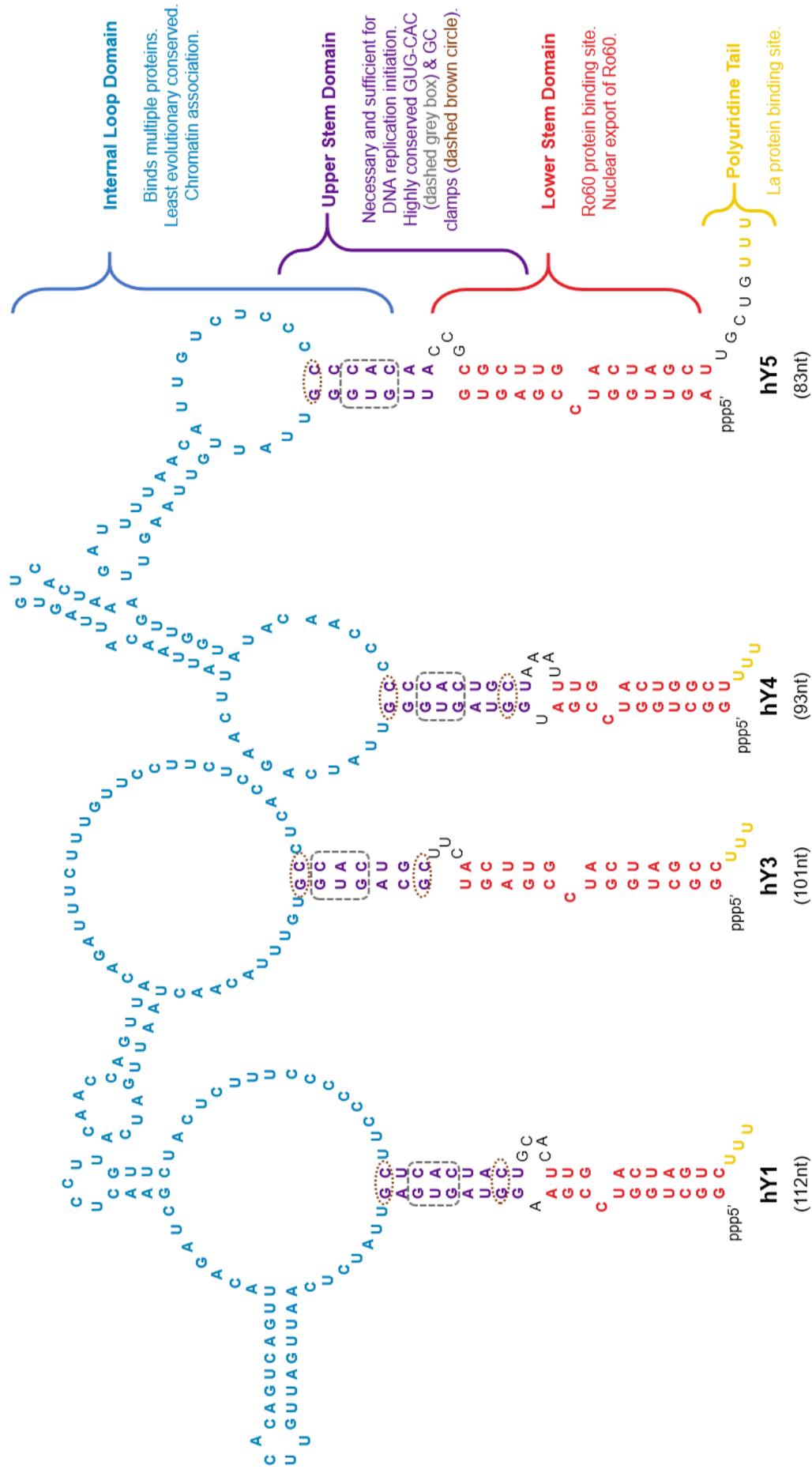
The Y RNAs were initially found as part of ribonucleoprotein (RNPs) complexes, where they were predominantly bound to Ro and La proteins in mammalian cytoplasm (249,251) but were later discovered to play an essential part in DNA replication initiation (127). Y RNA derived fragments have also been identified and increasing evidence suggests that they play a role in both healthy and diseased cells (252,253).

Y RNAs appear to play a diverse role in cellular function and communication (126,254)). Y RNAs and their fragments are present in a variety of tested cell types, cellular components (ie nucleus, cytoplasm, extracellular vesicles) and biofluids in both healthy and unhealthy human cells (252,253,255,256). As a result, they have been implicated in various diseases, including cancer (252), immunopathological diseases (257) and Systemic Lupus Erythematosus (250) and Sjögren's syndrome (258) and are being flagged as biomarkers. However, they have also been found to act in a beneficial manner in some diseases, including in the reduction of cardiac hypertrophy and renal injury and defence against influenza infection (259,260).

Y RNAs are highly conserved and are functionally redundant in vertebrates. Between 1 and 4 Y RNAs have been identified in all investigated vertebrate species which demonstrates gene duplication and loss during vertebrate evolution (261–264). These Y RNAs are generally found clustered on the same chromosome, often located between the EZH2 and PDIA4 genes, and possess their own transcriptional promoters (262,264). Distances between and order in which the different Y RNA genes are found in the genome tend to be well-conserved between species (261,263,264).

Humans have 4 functionally redundant Y RNAs (hY1, hY3, hY4 and hY5) (Fig1.8); their genes are located in a cluster on chromosome 7q36 (265,266) and are transcribed by RNA polymerase III (249). The transcribed Y RNAs are relatively small at 80-120 nucleotides in length and adopt a characteristic stem-loop secondary structure (267,268).

The stem-loop structure of Y RNAs results from the hybridisation of their 5' and 3' ends, resulting in the formation of highly conserved upper and lower stem domains with an internal loop and 3' polyuridine tail (Fig1.8) (126,267,268).



**Figure 1.8:** The characteristic stem-loop structure and sizes (nucleotides (nt)) of the four human Y RNAs (hY). The structural domains and associated functions are indicated: the variable internal loop (interacts with euchromatin and other proteins); the upper stem domain (essential for DNA replication) where the key GUG-CAC (grey dashed box) motif and GC clamps (brown dashed circles) are highlighted; the lower stem loop (Ro60 protein binding site); and the 3' polyuridine tail (La protein binding site)

The lower stem domain consists of a 7 base pair helix, containing a single nucleotide bulge on one strand and a 3-nucleotide bulge on the other, (269,270) which creates a distortion in the helix (271). This lower stem is the binding site for the highly conserved Ro60 protein (126) and when bound form Ro-RNPs, which have been implicated in ncRNA quality control, RNA stability and cellular responses to stress (particularly UV sensitivity) (272,273).

Ro60 possesses a toroidal shape with a positively charged central channel (which accommodates ssRNA) (274) and sequesters a potentially wide range of misfolded ncRNAs including 5S rRNA and snRNAs (275–277). It has two known binding sites; one on the surface, where either Y RNAs or the abnormal helix of a misfolded RNA might bind, and the central channel, where it is conjectured that a single strand 3' extension (commonly seen in pre-5S rRNA) of an already bound misfolded RNA may bind (278,279).

Y RNAs are believed to regulate the access of misfolded ncRNAs to Ro by sterically blocking misfolded RNA binding. Once a misfolded RNA has replaced the Y RNA, the mechanism that ultimately results in the removal (by refolding or degradation) of the misfolded RNA remains unknown (274,278,279). Y RNAs are also required for the translocation of Ro60 from the nucleus to the cytoplasm (273,280).

The 3' polyuridine tail is retained in Y RNAs, unlike many other mature RNAs. As a result, the La protein is able to bind to this Y RNA polyuridine tail to form La-RNPs (281), which are required for the successful termination of transcription by RNA polymerase III (282). La has been implicated in the nuclear retention of Y RNAs and identified as a potential candidate involved in the transportation of RNA polymerase III transcripts from the nucleus to the cytoplasm (281). However, Y RNA binding to La has been found to inhibit this chaperoning activity (283). It has also been suggested that the La protein protects Y RNAs from exonucleolytic degradation (281).

It is important to note that while the interactions of Y RNAs with Ro60 and La have been long documented, less than 50% of Y RNAs are bound to these proteins in human cell extracts. This implies that Y RNAs interact with additional protein complexes (126,284).

The internal loop domains are by far the least evolutionarily conserved domain of the Y RNAs. There is great diversity between the structures and primary sequences of the individual Y RNAs, which allows the loop domain to interact with a variety of different proteins (267,285) and may facilitate the specialisation of individual Y RNAs (254). In particular, the loop domains have been implicated in the coordination of Y RNA association with euchromatin (286).

The pyrimidine-rich regions of the loop domains of hY1 and hY3 are preferentially bound by nucleolin (284,287). Furthermore, polyprimidine tract-binding protein and hnRNP K also bind

with hY1 and hY3 but it is believed that this interaction requires the presence of La (254,287,288), which demonstrates the ability of Y RNAs to bind with more than one protein at any given time (287). As with La, it is believed that Y RNA inhibit the activity of these proteins (283). In addition, proteins such as RoRNP binding protein I preferentially interact with hY5 (289).

The interaction of Y RNAs with the range of loop-binding protein, alongside the Ro60 and La protein indicate that Y RNAs play an important role in a number of cellular processes (126). In the 2000s, Y RNAs were identified as an essential factor in human DNA replication initiation (127). Further analysis showed that the DNA replication activity was not disrupted by the removal of Ro60, La (or their binding sites) and nucleolin, suggesting that the crucial domain was not in the lower stem, internal loop or polyuridine tail (284,290).

It was later established that this vital DNA replication activity arises from the upper stem domain of Y RNAs (290).

The upper stem domain is a highly conserved, short double stranded nucleotide sequence that forms a locally destabilised  $\alpha$ -helix, containing a highly conserved GUG-CAC motif (290). This central motif is flanked by conserved G-C base pair, known as GC clamps, which in turn stabilise the destabilised helix (291).

Y RNAs were first recognized as key factors in DNA replication by the Krude lab, where they were identified from fractionated cytosolic extract tested on the *in vitro* DNA replication cell-free system (127). The removal of Y RNAs, via specific oligonucleotide degradation from cytosol, resulted in a reduction in actively replicating nuclei (ie DNA replication activity) in the human cell-free system (section 1.4). Further studies have found that Y RNAs are required for the initiation, but not elongation, of DNA replication (292).

Systematic mutagenesis studies have established that the upper stem domain of Y RNAs is essential and sufficient to function in DNA replication initiation. The highly conserved nature of the upper stem would explain Y RNAs' functional redundancy in DNA replication (290). The conserved GUG-CAC motif is of particular significance in Y RNAs' role in DNA replication initiation, as mutations of this motif abolish the replication activity observed in the human cell-free system (290,291). By contrast, complete removal of the lower stem and the polyuridine tail has no effect on DNA replication in the cell free system (290). The internal loop is thought to modulate Y RNA association with chromatin and although its removal does not affect replication activity, it does result in indiscriminate Y RNA binding to chromatin (286).

Similar observations have been seen in cell culture, whereby the targeted disruption of Y RNAs and addition of a synthetic hY1 upper stem RNA in living cells resulted in the inhibition

of DNA replication and cell proliferation and then restoration of DNA replication and cell proliferation, respectively (127,290,293,294).

Despite playing such a key role, the precise mechanism by which Y RNAs bring about DNA replication initiation remains unelucidated (126). It is highly probable that this mechanism involves the interaction of Y RNAs with various proteins/protein complexes.

Interaction of Y RNAs with replication initiation proteins including ORC, Cdc6 and Cdt1 have been determined (286,294). In addition, fluorescently labelled hY RNAs have been found to colocalise with ORC, MCM2, Cdt1 and Cdc45 on unreplicated chromatin during late G1, prior to DNA replication initiation. Once DNA replication elongation takes place, this association is no longer present, which is consistent with the observation that hY RNAs do not biochemically interact with the DNA replication fork proteins, MCM2-7 and GINS complexes, primase or any DNA polymerase (286).

Further analysis has shown that Y RNAs are predominantly present in the cytoplasm (254,295,296), although there remain conflicting accounts about the distributions of Y RNAs in the nucleus and cytoplasm. These can, in part, be explained by a bias generated based on the different experimental methodologies employed (286,297).

In colocalisation studies, it has been shown that Y RNAs are associated with euchromatin throughout the cell cycle (298); hY1, hY3 and hY4 mainly colocalise with one another at euchromatin whereas hY5 is predominantly localised to nucleoli (which further confounds the theory that it may be involved in rRNA biogenesis) (286,299).

Levels of chromatin-associated Y RNAs are higher during S-phase than either G1-phase or mitosis and it has been shown that Y RNAs are absent at actively replicating foci but quickly reacquire themselves with euchromatin during mid/late S-phase (298); an observation consistent with ORC binding (300). These chromatin colocalisation behaviours are also consistent with the observed associations of Y RNAs with various DNA replication proteins (298).

With this information in mind, it is clear that Y RNAs play an important role in DNA replication. However, the mechanisms involved and the proteins interactions that take place to bring about Y RNA-dependent DNA replication, remain elusive (126).

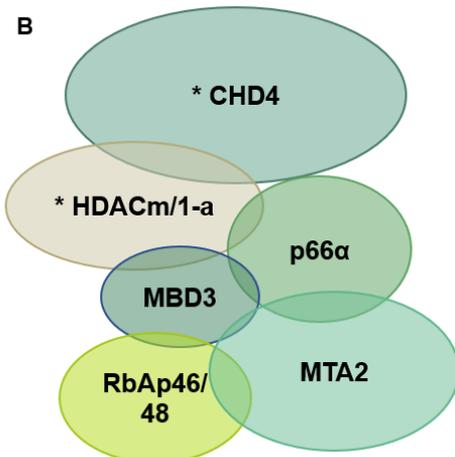
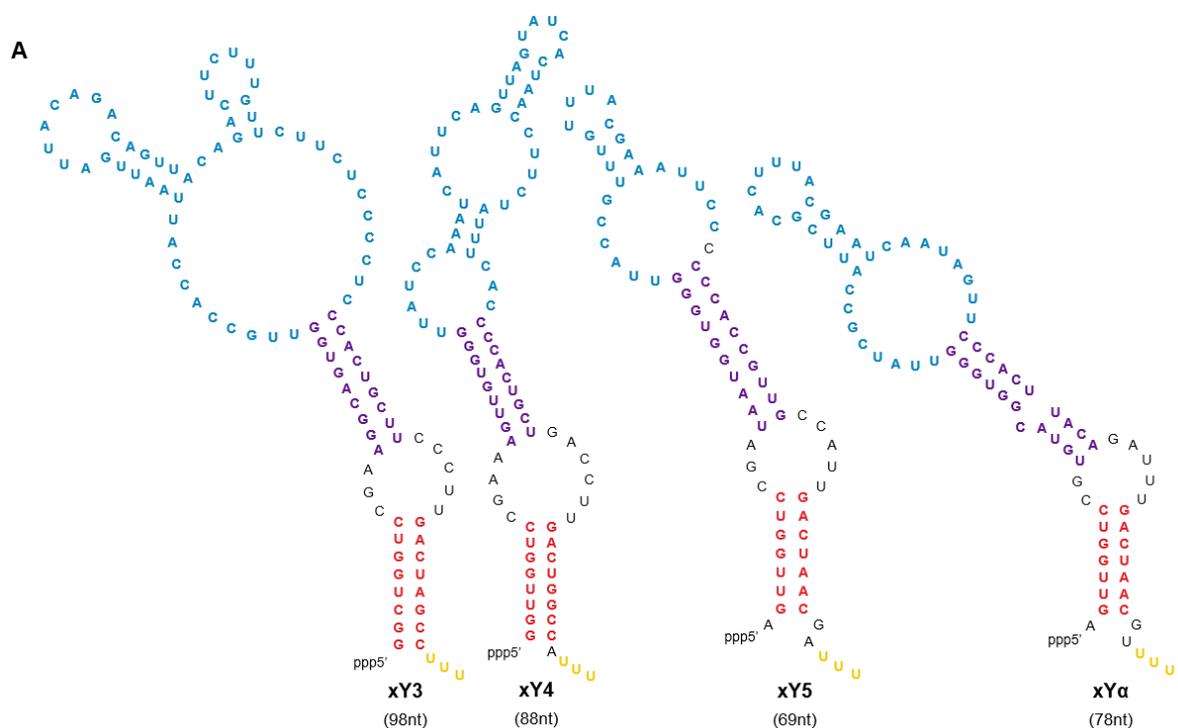
As with many other highly conserved RNAs, Y RNA homologues/orthologues have been identified in both eukaryotic (262) and prokaryotic model organisms (301–303). In addition to humans, Y RNAs have been identified in vertebrate species including Chinese hamster (304), *X. laevis*, *D. rerio* (294), *M. musculus* and *G. gallus* (262). A class of ncRNA, stem-bulge RNAs (sbRNAs), which possess a homologous structure with Y RNA, has been

identified in further eukaryotic model organisms, such as *C. elegans* (305), *D. melanogaster*, *B. mori* (306,307), *A. gambiae* (308) and *B. floridae* (262). Both these Y RNA homologues and sbRNAs found in *C. elegans* and *D. melanogaster* can substitute for hY RNAs in the human cell-free system, which highlights an evolutionarily conserved role for Y RNAs in DNA replication (290,294,304,305).

Y RNA orthologues have also been identified in prokaryotes, such as *S. enterica*, *D. radiodurans* and *M. smegmatis*, but these ncRNAs are not homologous with the vertebrate Y RNAs and do not substitute for vertebrate Y RNAs in DNA replication (301–303,308). Interestingly, some evidence may suggest that the most conserved bacterial Y RNA subclass may consist of a domain that mimics tRNAs (309).

## 1.7 Early *Xenopus laevis* Embryo DNA Replication and *X. laevis* Nucleosome Remodelling and Histone Deacetylation (xNuRD) complex

Like humans, *X. laevis* have 4 distinct Y RNAs (xY3, xY4, xY5 and xY $\alpha$ ) which are 60 – 100 nucleotides long and adopt the characteristic stem loop Y RNA structure (Fig1.9A) (290,294). These xY RNAs can functionally substitute for hY RNAs as DNA replication initiation factors in the human cell-free system (290). An essential *in vivo* role for xY RNA has been established in DNA replication during early *Xenopus* development; xY RNAs were found to be essential for DNA replication in *X. laevis* embryos following the Mid-Blastula Transition (MBT) of development (294).



**Figure 1.9:** (A) The characteristic stem-loop structure and sizes (nucleotides (nt)) of the four *X. laevis* Y RNAs (xY). The structural domains are indicated: the internal loop (blue), the upper stem (purple), the lower stem (red) and the 3' polyuridine tail (yellow). (B) The overall conformation of the Y RNA-independent DNA replication initiation factor, xNuRD (isolated from activated *X. laevis* egg extracts). The protein isoforms for the 6 subunits are CHD4, HDACm(1-a), MBD3, MTA2, p66 $\alpha$  and RbAp46/48 (RBBP7/4). \* indicates the enzymatic subunits.

As DNA replication is such a vital process in organism longevity, the protein machinery involved are highly conserved between species. The mechanism by which DNA replication initiation is carried out in *X. laevis* is similar to that of other vertebrate model organisms (15). However, in *X. laevis*, a substantial difference exists in the control of DNA replication during early embryo development (310).

During *X. laevis* development, there is an early developmental stage, known as the MBT, where there is a substantial shift in the control of DNA replication and other nuclear processes such as transcription. Prior to MBT, the embryo undergoes 12 rapid, synchronous cell divisions (59). During these divisions, S- and M- phases overlap and no growth phases take place (311). Ultimately, across nearly 8 hours, this leads to the formation of approximately 4000 smaller cells that occupy the same volume as the previously one large cell (59).

During pre-MBT, transcription does not take place and there is non-specific DNA replication origin firing, with a negligible amount of sequence specificity (311,312). The origin-binding ORC associates with sites spaced ~5-15Kb apart throughout the genome (313). In order to accommodate the rapid cell divisions and the overlapping of S- and M- phases, laminal nuclear membranes form around individual chromosomes during anaphase. These separate structural units, called karyomeres, requires normal segregation of mitotic spindles and chromosome formation, but karyomere formation is independent of DNA replication (311).

The xY RNAs are not required for DNA replication throughout pre-MBT, as xY RNA-depleted embryos developed normally until they reached MBT, at which point DNA replication was stalled and the embryos perished (294). In collaboration with the Smith (Francis Crick Institute) and Sale (MRC-LMB, Cambridge) groups, the Krude group has identified the factor that facilitates Y RNA-independent replication in early *X.laevis* embryos (pre-MBT) as a specific isoform composition of the chromatin remodelling complex Nucleosome remodelling and histone deacetylation complex (Fig1.9B) (239).

The highly evolutionarily conserved, NuRD is a six-subunit protein complex that possesses both ATP-dependent chromatin remodelling and histone deacetylase activities (314–318). Although NuRD is one of four ATP-dependent chromatin remodellers, it is unique in exhibiting its second, histone deacetylase, enzymatic activity (44). These combined activities enable NuRD to regulate overall chromatin structure (318,319).

NuRD has been implicated in multiple cellular processes, including gene expression/transcription (involved in transcriptional silencing), DNA damage repair, chromatin assembly and cell cycle progression (318,319). Until the identification of xNuRD as a DNA replication initiation factor, there was no known role for a NuRD complex in DNA

replication (239). However, depletion of the xNuRD subunit, CHD4, had been shown to result in delayed S-phase progression and defects in pericentric heterochromatin and its assembly, implying a role in chromatin structure and cell cycle progression (37).

The six NuRD subunits have multiple (often mutually exclusive) isoforms, which facilitate the formation of various different complex isoforms, which are often cell type specific and in response to physiological signals (315,318,320). NuRD is comprised of the two enzymatic subunits, CHD3/4 (ATP-dependent nucleosome remodeller with helicase activity) and HDAC1/2 (histone deacetylase), and the 4 non-enzymatic subunits, MTA1/2/3, MBD2/3, RbAp46/48 (aka RBBP7/4) and p66 $\alpha$ / $\beta$  (318,321).

This xNuRD, involved in DNA replication initiation, is comprised of a specific assortment of the subunit isoforms; the subunits are CHD4, MTA2, RBBP7/RbAp46, MBD3, p66 $\alpha$  and a class 1 HDAC, HDACm (and/or HDAC1) (239). The maternally deposited HDACm (aka HDAC1-a) accumulates during oogenesis, but dissociates from xNuRD and, along with p66 $\alpha$ , reduces in abundance in the cell, following MBT. It has been conjectured that this HDACm/HDAC1-a is the subunit that directs the specialisation of xNuRD to DNA replication initiation, although further analysis must be undertaken to assess this hypothesis. Notably, NuRD isolated from human immortalised cell lines do not contain this HDACm and are not able to substitute for Y RNAs in the human cell-free system (239,322,323).

What remains unclear about xNuRD is how it brings about DNA replication initiation. As the DNA replication landscape of pre-MBT displays a highly delocalised and dynamic array of replication origin sites, it is possible that xNuRD may play a role in this reduced origin specificity (312). Whether xNuRD brings about DNA replication in a site specific or unspecific fashion has yet to be elucidated.

When *X. laevis* embryo development reaches MBT, xNuRD is no longer essential for DNA replication initiation, and Y RNAs take up their role of vital DNA replication initiation factors (239,294). Post-MBT the *X. laevis* embryos establish a typical cell cycle, where cell divisions become asynchronous and slower with distinct G1 and G2 gap phases; S- and M- phases no longer overlap (310). Furthermore, more specific DNA replication origins are established and the laminal nuclear membranes, karyomeres, are no longer present (311). Finally, zygotic transcription begins, and cell motility is observed (324).

The roles of Y RNAs and the functionally equivalent, xNuRD are crucial in the replication of genomic DNA. However, the mechanism by which they bring about DNA replication initiation remains elusive. How these factors impact the specification of replication origins is also unknown. The development and enhancement of the NGS method, iniSeq for origin

identification, offers a promising avenue for the elucidation of the roles Y RNAs and xNuRD play.

## **Chapter 2: Aims and Objectives**

The aim of my PhD is to explore and evaluate the characteristics that influence the specification and activation of human DNA replication origins; and further, to elucidate the impact of two previously identified initiation factors, the non-coding hY RNAs and the chromatin remodelling complex xNuRD on the specification and activation of human DNA replication origins. I also aim to establish whether the Y RNA-interacting chromatin remodelling complex Polycomb repressive complex 2 (PRC2) functions as a DNA replication factor via its interaction with hY RNAs.

The first objective of my PhD is to develop and improve the iniSeq method for the identification of human DNA replication origins, to establish a novel and valid “density-substitution initiation-site sequencing” (ds-iniSeq) technique. This method and subsequent bioinformatical analyses should be capable of reliably identifying the locations and size of human DNA replication origins (ie origin specification). It should improve significantly on the iniSeq method, by determining the relative extent of DNA replication at the identified origins (ie origin activation efficiency).

The second objective is to perform in-depth genome-wide analysis of the locations and activities of DNA replication origins identified by ds-iniSeq, focussing on the impact of genomic and epigenetic features on origin specification and activation.

The third objective is the analysis of the function of established DNA replication initiation factors for origin specification and activation; the key advantage of ds-iniSeq over other established methods for mapping DNA replication origins is the ability to successfully biochemically manipulate the *in vitro* DNA replication reaction, from which all DNA (both replicated and unreplicated) is obtained. This will enable me to address the following questions:

- 1 – Does the absence or presence of non-coding hY RNAs affect DNA replication origin specification, activation and efficiency?
- 2 – Does xNuRD (which can functionally substitute for Y RNAs during early amphibian development) affect DNA replication origin specification, activation and efficiency in the absence of Y RNAs?

The fourth objective is to analyse further elongation of DNA replication forks following the activation of replication origins *in vitro*, using a relative site/origin activity. The successful adaptation of ds-iniSeq to “density-substitution elongation-site sequencing” (ds-eloSeq) should enable me to assess DNA replication elongation genome-wide over an extended replication reaction time period. The ultimate aim is to use ds-eloSeq to assess the effects of Y RNA

removal and xNuRD addition on DNA replication elongation following replication origin activation *in vitro*.

The final objective is the functional analysis of the polycomb repressive complex 2 (PRC2) as a Y RNA-binding DNA replication initiation factor. From mass spectrometric data generated by M. Kowalski (PhD student *circa* 2014 - Krude group), I identified all the core subunits of PRC2 as proteins that possess DNA replication activity (in the human cell-free system) and bind to hY RNAs. I will perform a biochemical investigation of PRC2 to determine if it plays a role in DNA replication initiation in the human cell-free system and consequently provides a mechanistic link between Y RNAs, chromatin remodelling and DNA replication in human cells.

## **Chapter 3: Materials and Methods**

### **3.1 Cell culture**

Human HeLa S3 cervical carcinoma and human EJ30 bladder carcinoma cells were cultured as proliferating monolayers on 145cm<sup>2</sup> plastic plates (Nunc). All cells were grown in Dulbecco's Modified Eagle Medium (DMEM), 10% Foetal Calf Serum, 1% Penicillin/Streptomycin (Gibco), at 37°C, 10% CO<sub>2</sub>.

### **3.2 Cell Propagation**

Cells were maintained by splitting 1:4 to 1:6 every 3-5 days, when they were fully confluent. The media was removed from the cells, which were washed in Dulbecco's phosphate buffer saline (PBS). After the PBS was removed, the cells were trypsinised (Trypsin (Invitrogen) 1:6 dilution (in PBS)). Once fully trypsinised, pre-warmed media was added, and the solution of cells were separated between new plates containing prewarmed media.

### **3.3 Preparation of Human cytosolic extract, S100**

Cytosolic HeLa extract (CilBioTech) was centrifuged at 100,000rcf for 1 hour, to produce HeLa S100 cytosol (human cytosol). The S100 was snap frozen as beads and stored at -80°C. The S100 was then defrosted on ice, when required.

The protein concentration of the S100 was determined using a Bradford assay (BioRad) in accordance with the manufacturer's instructions. Absorbances at 595nm were compared to a standard curve of known BSA protein concentrations.

### **3.4 Preparation of EJ30 late G1- and S- phase template nuclei**

#### *3.4.1 Late G1-phase synchronisation*

Synchronised template nuclei were isolated from mimosine-arrested, human bladder carcinoma, EJ30 cells. This preparation was performed in accordance with Krude (1), and all references therein.

Prior to harvesting, EJ30 cells were synchronised and arrested in late G1-phase. EJ30 cells at ~60% confluency were incubated with 0.6mM mimosine (Sigma) for 24 hours. The cells were then subjected to the nuclei isolation preparation (see below).

#### *3.4.2 S-phase synchronisation*

Synchronised template nuclei were isolated from Thymidine-arrested, human bladder carcinoma, EJ30 cells. This preparation was performed in accordance with Krude (2), and all references therein.

Prior to harvesting, EJ30 cells were synchronised and arrested in S-phase. EJ30 cells at ~60% confluency were incubated with 2.5mM thymidine (Sigma) for 24 hours. The thymidine was removed and cells released into fresh pre-warmed media for 30 minutes. The cells were then subjected to nuclei isolation preparation (see below).

#### *3.4.3 Nuclei isolation preparation*

The following preparation was performed at 4°C. The synchronous EJ30 were washed once with ice-cold hypotonic buffer and then incubated in the same buffer for approximately 5 minutes. The hypotonic buffer was discarded, cells were removed using a scraper and underwent dounce homogenisation (Duchan dounce homogeniser with loose-fitting pestle). The resultant lysate underwent centrifugation for 3 minutes at 2,300rcf.

The supernatant was discarded and the pellet was resuspended and washed 2-3 times in SuNaSpBSA (250 mM sucrose, 75 mM NaCl, 0.5 mM spermine trihydrochloride, 0.15 mM spermidine tetrahydrochloride, 3% bovine serum albumin). Following the final wash, the pellet of nuclei was resuspended in a residual volume of SuNaSpBSA, aliquoted and snap frozen. The nuclei preparations were stored at -80°C for up to 3/4 months.

### **3.5 *In Vitro* DNA replication experiment using the Human cell-free system**

#### *3.5.1 Preparation of the replication reactions:*

The human cell-free system was used to measure DNA replication activity and performed in accordance with Krude (1) and Christov *et al* (3) and references therein.

Template G1 or S-phase nuclei (preparation in section 3.4) were incubated (37°C) for 3 hours in the presence of ~100µg human cytosol (S100; preparation in section 3.3), a mixture of NTPs and dNTPs (including digoxigenin-11-dUTP (Roche) and a lower amount of dTTP), and the ATP-regenerator, creatine kinase/phospho-creatine. The reaction was made up to a total volume of 50µl with a physiological replication buffer (100mM K-Ac, 20mM K-Hepes

(pH7.8), 1M DTT, 0.5mM EGTA). Prior to this 3-hour incubation, the constituents of the replication reaction were pipetted on ice.

This reaction enables the template DNA to undergo replication initiation and digoxigenin-dUTP to be incorporated into the newly synthesised DNA.

### *3.5.2 Manipulation of the replication reactions:*

This *in vitro* replication experiment allows for manipulation of the reactions through the removal and or addition of factors, to test their effect on DNA replication activity.

The manipulations used in this thesis include:

1. DNA oligonucleotide depletion human cytosolic S100 extract.
  - Mock depletion of S100 whereby a T3 DNA oligonucleotide (0.4 $\mu$ M) was incubated with S100 (used to control for the addition of an oligonucleotide to the replication reaction) for 1 hour and 20 mins (RT) prior to its addition to the human cell-free system.
  - hY RNA depletion of S100 whereby hY RNA specific DNA oligonucleotides (0.2 $\mu$ M hY1, 0.1 $\mu$ M hY3, 0.05 $\mu$ M hY4, 0.05 $\mu$ M hY5) were incubated with S100 for 1 hour and 20 mins (RT) prior to its addition to the human cell-free system.

2. The addition of xNuRD

Partially fractionated xNuRD (~40 $\mu$ g) (see section 3.6) was added to a replication reaction following the removal of hY RNAs (by oligonucleotide depletion).

3. The chemical inhibition of PRC2 in S100 cytosolic extract

In order to inhibit the PRC2 complex, the chemical inhibitor UNC1999 (0.1, 0.3, 1, 3, 10, 30 and 100 $\mu$ M) was incubated with the S100 for 30 mins (RT) prior to its addition to the cell-free system. DMSO was used as a control, whereby the same volume of DMSO was incubated with the S100 for 30 mins (RT) prior to its addition to the cell-free system.

4. Antibody inhibition of PRC2 in S100 cytosolic extract

Antibodies for SUZ12 (ab175187) and EED (ab4469) were incubated with S100 (concentrations of 1:50, 1:100 and 1:200) for 1 hour (RT) prior to its addition to the cell-free system. BSA was used as a control and was incubated with S100 (concentrations of 1:50, 1:100 and 1:200) for 1 hour (RT) prior to its addition to the cell-free system.

#### 5. Chemical and antibody inhibition of PCR2 in S100 cytosolic extract

Cytosolic extracts were incubated with; DMSO and EED or SUZ12 antibodies or BSA, or 3 $\mu$ M UNC1999 and EED or SUZ12 antibodies or BSA. The antibodies were at concentrations of 1:50 and 1:100. The antibodies/BSA were incubated with the S100 for 30 mins (RT) and then the UNC1999/DMSO was added to the S100, which was then incubated for a further 30 mins (RT). The resultant S100 was then used in the appropriate replication reactions.

#### 6. Immunodepletion of SUZ12 and EED in S100 cytosolic extract

Immunodepleted cytosolic extract (see section 3.11) was substituted for untreated cytosolic extract in the human cell-free system.

#### *3.5.3 Stopping the replication reaction and slide preparation and analysis:*

The nuclei are fixed with 4% paraformaldehyde and adhered to polylysine-coated coverslips by centrifugation (5 minutes; 1000rcf; RT) through 30% sucrose. The nuclei on the coverslips were incubated with propidium iodide (Sigma; stain for DNA) and fluorescein Fab anti-digoxigenin antibodies (Roche; stain for incorporated digoxigenin-dUTP and therefore newly replicated DNA). The coverslips were then washed and mounted onto glass slides.

The nuclei were imaged using confocal microscopy (Olympus FV3000; 30X lens magnification; 488nm and 561nm lasers), and replicating and non-replicating nuclei were counted, in order to determine the percentage of replicating nuclei. Where possible, the means and standard errors of the percentage of replicating nuclei underwent statistical analysis by a 2-tailed student T-test (unequal variance) at the 5% significance level, or ANOVA and Tukey's post-hoc tests.

The resultant images of nuclei were also analysed for the intensity of fluorescence (integrated density) of the digoxigenin incorporation in each nucleus, using a FIJI script (GG

FIJI script) written by Guillaume Guilbaud (LMB, Cambridge). Nuclei images were visualised, and scale bars were inserted using FIJI.

### **3.6 Preparation of xNuRD from *Xenopus laevis* eggs**

Crude *X. laevis* egg extracts were prepared and provided by K.S Dingwell (J.C Smith lab group, Francis Crick Institute, London) (4). The crude extract was treated with xY RNA antisense oligonucleotides (1.7 $\mu$ M xY3, 0.7 $\mu$ M xY4, 0.5 $\mu$ M xY5, 0.1 $\mu$ M xY $\alpha$  (Sigma)) for 1 hour at room temperature (RT).

The crude extract underwent the following fractionation steps, to produce partially purified xNuRD:

#### *3.6.1 Ammonium sulphate precipitation:*

The extracts underwent 20%, 45% and 100% ammonium sulphate precipitations. The appropriate mass of ammonium sulphate (Sigma) (0.106g/1ml for 20%; 0.146g/1ml for 45%; 0.383g/1ml for 100%) was dissolved in the extracts, incubated at 4°C for 1 hour and centrifuged for 30 minutes at 4°C, 16,100rcf. The supernatant was used in the next precipitation step and the pellet was re-suspended in 500 $\mu$ l replication buffer. The pellet containing samples and the final supernatants were snap frozen and stored at -20°C.

#### *3.6.2 Sucrose gradient ultracentrifugation:*

The 20% - 45% ammonium sulphate fraction underwent further fractionation using sucrose gradient ultracentrifugation. A 40% to 15% sucrose gradient (in replication buffer) was prepared and left at 4°C for 20 minutes. An aliquot of 500 $\mu$ l of the ammonium sulphate fraction was loaded on to each gradient column and underwent centrifugation at 34,000rpm (MLS-50 rotor, Beckman), 0 psi for 18 hours (4°C). Following centrifugation, fractions of 500 $\mu$ l were collected, aliquoted and snap frozen and stored at -20°C.

#### *3.6.3 Bradford (BioRad) assay:*

The protein concentration of xNuRD fractions were determined using a Bradford assay (BioRad) in accordance with the manufacturer's instructions. Absorbances at 595nm were compared to a standard curve of known BSA protein concentrations.

### 3.7 Density-substitution initiation-site sequencing (ds-iniSeq)

#### 3.7.1 Replication reactions to generate replicated DNA:

Ds-iniSeq is an adapted method of iniSeq and is used to determine human DNA replication origin sites. As with iniSeq, ds-iniSeq is based on an adaptation of the human cell-free system replication reaction (see section 3.5), whereby synchronised G1-phase nuclei are incubated (37°C) for 15 minutes, with ~300µg human cytosol from proliferating HeLa cells, a mixture of NTPs and dNTPs (3mM ATP; 0.1mM of GTP, CTP, UTP; 0.1mM of dATP, dGTP, and dCTP; and 0.1mM **BrdUTP** (replaced dTTP); 0.5 mM DTT (Sigma) in 40 mM K-HEPES pH 7.8 7 mM MgCl<sub>2</sub>), and the ATP-regenerator, creatine kinase/phospho-creatine.

The replication reactions can be manipulated to assess the effect of a given factor on human DNA replication origin specification and activation. In this thesis, ds-iniSeq was used to assess the effect Y RNA depletion and xNuRD addition.

Four experimental conditions were used: standard positive control reaction (unchanged human cytosol was added to the reaction); a mock depletion using a T3 oligonucleotide (to control for the addition of a DNA oligonucleotide to the replication system); a Y RNA depletion; and the depletion of Y RNAs + the addition of partially purified xNuRD. The mock depletion and Y RNA depletions were achieved by the incubation of the human cytosol with DNA oligonucleotides (0.4µM T3 oligonucleotide for mock; 0.2µM hY1, 0.1µM hY3, 0.05µM hY4, 0.05µM hY5 for Y RNA depletion) for 1 hour 20 minutes (RT) prior to use in the replication reaction.

Alongside the generation of DNA for sequencing, DNA replication experiments (using the sample experimental conditions as those in ds-iniSeq) were conducted in the cell-free system (see protocol above). Images of the 3-hour samples were analysed by both the GG FIIJ script and were counted to determine percentage of replicating nuclei

#### 3.7.2 Isolation of total genomic DNA:

Replication reactions were stopped by the addition of ice-cold PBS. They then underwent centrifugation (10 mins; 4°C; 16,100rcf), the supernatant was removed, and more PBS was added to the tubes. The nuclei from the 3 tubes (of the same experimental condition) were combined into 1 tube and spun again to recover the nuclei.

The supernatant was removed (to remove any excess reagents) and the pellet (nuclei) was resuspended in DNA purification buffer (10mM Tris-HCl pH8.0, 125mM NaCl, 1mM EDTA, 1% sodium lauroyl sarcosinate w/v, 0.01% SDS) + 2mg/ml proteinase K, and incubated for 18 hours (55°C).

Total genomic DNA was isolated by phenol/chloroform extraction and ethanol precipitation, resuspended in DNA buffer 2 (10mM Tris-HCl pH8.0, 125mM NaCl, 1mM EDTA) and fragmented by sonication (Covaris ME220) resulting in fragments of 100-1000 bp. Sonicated DNA was tested on a 2% agarose gel.

### *3.7.3 Separation of HL and LL DNA:*

The resultant fragmented DNA consisted of a mixture of Light-Light (LL) and Heavy- Light (HL) DNA and these were then separated on a caesium sulphate gradient.

The samples were loaded into Optiseal Pollyallomer tubes and caesium sulphate solution (in TE buffer (pH7.4); refractive index of 1.3700 - 1.3710) was added to bring the total volume to 5ml. These samples underwent ultracentrifugation at 55,000rpm (NVT90 rotor (Beckman)) for 22 hours (20°C). The gradient was fractionated using a peristaltic pump from the bottom of the tube. The first 6 drops were discarded; subsequently 5 drops were collected per fraction resulting in approximately 25 fractions.

Every other fraction was tested with a refractometer to provide refraction indices (HL DNA is present at RI = 1.3700 and LL DNA is present at RI = 1.3675). Fractions were analysed using nanodrop spectrophotometry to indicate relative DNA concentration.

From this 1<sup>st</sup> caesium sulphate gradient, 1-3 fractions containing the LL DNA were pooled and prepared for sequencing. In addition, 2-3 fractions containing the HL DNA were then loaded onto a 2<sup>nd</sup> caesium sulphate gradient and fractionated and analysed as before.

From this 2<sup>nd</sup> caesium sulphate gradient, 1-3 fractions containing the HL DNA were pooled and prepared for sequencing.

### *3.7.4 Preparation of HL and LL DNA for sequencing:*

The pooled HL and LL DNA samples were desalted using a PD MidiTrap (G-25) column (GE Healthcare) in accordance with the manufacturer's instructions for the gravity protocol (TE pH7.4 was used as the elution buffer).

The resultant HL and LL samples underwent an isopropanol precipitation (whereby samples were spun for 1 hour at 4°C) in the presence of 0.055 µg/µl RNA grade glycogen (Thermo Scientific). The resultant pellets were resuspended in 25µl DEPC H<sub>2</sub>O and analysed by Illumina sequencing at the Francis Crick Institute, London (xNuRD project samples) or the MRC-LMB, Cambridge (preliminary tests of the method).

### 3.7.5 Next-generation DNA sequencing:

All library production and subsequent sequencing was performed by our collaborators; J.C. Smith's Group at the Francis Crick Institute, or J. Sale's group at the MRC-LMB, Cambridge. For next-generation Illumina sequencing at the Francis Crick Institute, libraries were generated using the KAPA DNA HyperPrep preparation kit according to the manufacturer's instructions and the sequencing was performed using the Illumina HiSeq 4000 platform (75bp SR) according to the manufacturer's instructions.

For next-generation Illumina sequencing at the MRC-LMB, libraries were generated using the Illumina NEBNext sample preparation kit (New England Biolabs) according to the manufacturer's instructions and the sequencing was performed using the Hi-Seq 4000 platform according to the manufacturer's instructions.

### 3.8 Density-substitution elongation-site sequencing (ds-eloSeq)

Density-substitution elongation-site sequencing (ds-eloSeq) was developed from the ds-iniSeq method, to assess DNA replication elongation in then human cell-free system, and the impact of various factors on elongation. The method for ds-eloSeq was the same as that of ds-iniSeq (described in 3.8), with the exception that the *in vitro* replication reaction was carried out for 3 hours, to allow for elongation to take place.

In summary, BrdU was incorporated into newly synthesised/replicated DNA during a 3-hour reaction. Total DNA is then purified and fragmented into sizes of 100-1000bp. Replicated DNA (HL) was separated from unreplicated DNA (LL) through two rounds of caesium sulphate density gradients. The LL was isolated from the first gradient and the HL was isolated from the second gradients. The HL and LL were desalted and sent for Illumina NGS.

As with ds-iniSeq, the replication reactions were manipulated to assess the effect of a given factor on human DNA replication origin specification and activation. In this thesis ds-eloSeq was used to assess the effect Y RNA depletion and xNuRD addition.

The same experimental conditions were used for the ds-eloSeq xNuRD experiments as were in the ds-iniSeq xNuRD experiments: standard positive control reaction; a mock depletion using a T3 oligonucleotide; a Y RNA depletion; and the depletion of Y RNAs & the addition of partially purified xNuRD. The ds-eloSeq experiments were carried out alongside the corresponding ds-iniSeq experiments.

### 3.9 Bioinformatics

#### 3.9.1 Processing and alignment of sequenced reads – *ds-iniSeq* and *ds-eloSeq*:

Read quality of the sequencing data was assessed by FastQC. Where there were multiple files per reaction condition, the files were catenated together prior to alignment. The reads in the subsequent files were aligned to the hg38 human genome (GRCh38); PRC duplicates and non-unique reads were removed using Bowtie2, SAMtools and BEDtools, which resulted in .bam files for each experimental condition. The processing and alignment were carried out by T.Krude. The resultant files were used for bioinformatic analysis.

#### 3.9.2 Peak calling – *ds-iniSeq*:

To identify the DNA replication origins, the aligned reads for the 15-minute *ds-iniSeq* samples were analysed using SeqMonk (version 1.46.0) ChIP probe generation, where the HL (raw read numbers were normalised to the lowest number for each replicate set of data) sample acted as the ChIP sample and the LL DNA acted as the input DNA (fragment size = 200bp;  $p = 10^{-9}$ ), and a read count quantitation (normalised for per million reads). The called sites that were within 500bps of one another were grouped together. This generated a list of replication origin sites and corresponding origin firing efficiency values, which were subsequently used for further analysis. The R studio package, “*GenomicRanges*” (Bioconductor) was used to identify the common origins to all 3 replicates of each condition; the origins of replicate 1 which overlapped replicates 2 and 3 were used for analyses in results chapters 5 and 6.

#### 3.9.3 Assessment of elongation – *ds-iniSeq* and *ds-eloSeq*

Using the SeqMonk feature probe generator and read quantification tools, the relative activities of the sites 0.25Kb, 0.5Kb, 0.75Kb, 1Kb, 1.5Kb, 2Kb, 5Kb, 7.5Kb, 10Kb, 15Kb, 20Kb, 30Kb and 40Kb upstream and downstream of the *ds-iniSeq* origin locations in the *ds-iniSeq* and *ds-eloSeq* samples were calculated.

#### 3.9.4 Peak calling – Wilkes-Mookerjee (WM) method

To identify the replication origins, the aligned reads for the HL and LL DNA samples for the control 15-minute *ds-iniSeq* and the 3-hour *ds-eloSeq* samples were normalised to the total read count of each individual file. The normalised LL files were subtracted from the corresponding normalised HL files in 500bp windows. Origin sites were called where the normalised HL minus normalised LL read counts were greater than the threshold of 0.00003. These sites were quantified by calculating the relative read count of the normalised HL minus

normalised LL above the threshold. The script (appendix A3.1) was written by S.Mookerjee (St Catharine's College, Cambridge), in collaboration with me and executed in Python3.6.

### 3.9.5 Bioinformatical analysis:

Further analysis was undertaken using R studio (version 1.1.463). Overlap analyses were conducted using the “*GenomicRanges*” package (Bioconductor), where the minimum overlap was one nucleotide. Venn diagrams representing the overlaps were visualised using the eulerAPE 3.0.0 software. The data, including positions, was visualised on the Integrated Genomics Viewer (version 2.4.13).

Alternative origin calling methods were obtained from the following:

The original raw iniSeq files were realigned to the GRCh38 genome by T.Krude (Dept. of Zoology, Uni. Of Cambridge) (5). The SNS-seq data was conducted on EJ30 cells (fragment size 1Kb) (MRC-LMB, Cambridge) and were generated using the GRCh38 human genome (sites called by MACS) by L. Koch-Lerner and G. Guilbaud. The OK-seq (6) and bubble-seq (7) analysis were conducted on the GM06690 cell lines and were generated using the GRCh37 human genome.

The genomic features used for comparative analyses were obtained from the following sources:

Genes, TSS, promoters and enhancers were downloaded from the ensemble databases ([http://www.ensembl.org/Homo\\_sapiens/Info/Index](http://www.ensembl.org/Homo_sapiens/Info/Index)). CpG islands were downloaded from the SeqMonk internal database (version 1.46.0). The G4s were provided by G.Guilbaud (LMB-MRC, Cambridge). All of these data were generated in the GRCh38 genome.

The ChIP-seq data for the epigenetic features (HCT116 and PC3 cell lines) were obtained from the deposited data of the Encode project ([https://www.encodeproject.org/chip-seq-matrix/?type=Experiment&replicates.library.biosample.donor.organism.scientific\\_name=Homo%20sapiens&assay\\_title=Histone%20ChIP-seq&status=released](https://www.encodeproject.org/chip-seq-matrix/?type=Experiment&replicates.library.biosample.donor.organism.scientific_name=Homo%20sapiens&assay_title=Histone%20ChIP-seq&status=released)). The epigenetic features and their experiment codes were: H2A.Z (ENCFF459ZNJ), H3K4me1/2/3 (ENCFF874IUB/ENCFF786ORW/ENCFF610XEV), H3K9ac (HCT116; ENCFF278NDW, PC3; ENCFF645AQC), H3K27ac (HCT116; ENCFF560CAH, PC3; ENCFF890UYI), H3K79me2 (ENCFF680FOJ), H3K9me3 (HCT116; ENCFF327SRD, PC3; ENCFF513SGF), H3K27me3 (HCT116; ENCFF584YOT, PC3; ENCFF065HKC), H4K20me1 (ENCFF826BJO) and H3K36me3 (ENCFF959DKV). All but the H3K79me2 data were generated from the GRCh38 genome.

The replication timing data was generated by repli-seq analysis conducted in HeLa cells (8).

### 3.9.6 Realignment of 3<sup>rd</sup> party genomic data

The H3K79me2 CHIP-seq, OK-seq and bubble-seq data were originally aligned to the hg19 (GRCh37) human genome and the repli-Seq data was originally aligned to the hg18 (NCBI36) human genome.

These data were lifted over to the GRCh38 human genome using the online UCSC lift over tool (<https://genome.ucsc.edu/cgi-bin/hgLiftOver>).

### 3.10 Western blot:

A standard Western blot protocol was used to determine the presence of specific proteins of interest.

Samples were treated with 4X SDS loading buffer (200mM Tris HCl pH6.8, 4% w/v SDS, 20% w/v glycerol, 200mM DTT, ≤0.2% w/v bromophenol blue) (95°C for 10 minutes) and run on appropriate percentage (8%-10%) polyacrylamide SDS gels in 1X SDS running buffer. A low molecular weight protein marker (BioRad) was run alongside the samples.

Transfer to a nitrocellulose membrane (Scientific Laboratory Supplies) was carried out for 1 hour at 25V, 400mA and then checked with 0.1% Ponceau S (Sigma), 5% Acetic acid.

Following destaining (PBS/0.1% Tween 20), the membrane was blocked with 5% milk (PBS/0.1% Tween 20) for 1 hour (RT) or overnight (4°C). The membrane then underwent the primary antibody incubation for 2 hours (RT) or overnight (4°C), with the appropriate antibody ( $\alpha$ -EED (abcam ab4469) 1:200;  $\alpha$ -SUZ12 (abcam ab175187) 1:1000) in 5% milk (PBS/0.1% Tween 20).

The membrane was subsequently washed with 5% milk (PBS/0.1% Tween 20) for 2 x 20mins (RT) and then underwent the secondary antibody incubation (1:1000 anti-rabbit or anti-mouse (in PBS/0.1% Tween 20) was used for the appropriate primary antibody) for 1.5 hours (RT).

Finally, the membrane was washed in PBS/0.1% Tween 20 for at least 30 mins, rinsed in PBS and developed with the ECL prime Western developing kit (GE Healthcare – Amersham).

### **3.11 Immunodepletion of human cytosolic extract:**

The PCR2 subunit proteins, SUZ12 and EED were removed from human cytosolic extract (S100) by immunodepletion with Protein G Dynabeads (Invitrogen) in accordance with the manufacturer's instructions.

The storage solution of magnetic protein G Dynabeads (50µl per condition) was removed and the beads were resuspended in either 200µl PBS/0.02% Tween 20 (for the "beads only" control) or 200µl PBS/0.02% Tween 20 and 10µg of the appropriate antibody (α-EED (ab44469), α-SUZ12 (ab175187)). The samples were incubated with rotation (5rpm) for 45 mins (RT).

The supernatant was removed and the beads-antibody complex was washed with PBS/0.02% Tween 20. To this complex, 100µl of S100 cytosolic extract (14 µg/µl) was added and the samples were incubated with rotation (5rpm) for 45 mins (RT). The cytosolic extract (supernatant) was removed and kept for later use. The bead-antibody-protein complex was washed with 200µl PBS/0.02% Tween 20, transferred to a new tube and resuspended in 50µl PBS.

For the immunodepletion of SUZ12, the cytosolic extract underwent the above immunodepletion protocol twice. All immunodepleted samples were snap frozen and stored at -20°C until their use in the human cell-free system and assessment by Western blot.

### **3.12 Agarose gels:**

Agarose gels were used to act to assess the presence and native size of DNA or RNA in quality control of iniSeq DNA and total RNA preparations. Orange G nucleotide was added to each sample. Samples and a low range molecular weight marker (RiboRuler) were loaded onto a 2% agarose gel (0.5X TE buffer) in the presence of ethidium bromide. The gels were run in 0.5X TE running buffer at 70V for up to 1.5 hours and imaged using by UVP software.

### **3.13 Total RNA preparation:**

A Trizol (Invitrogen) extraction protocol for cells grown in a monolayer was carried out in accordance with manufacturer's instructions in order to extract total RNA from untreated and mimosine treated (24-hour mimosine treatment) EJ30 cells.

Untreated and mimosine treated cells were washed with PBS and scraped off the culture plates and transferred into an Eppendorf tube. Trizol (0.57ml per 0.25ml of cells) was added

and incubated for 5 minutes (RT). Following which, chloroform (0.2ml per 1ml of Trizol used) was added and incubated for 2-3 minutes and the samples were centrifuged for 15 minutes at 12,000rcf (4°). The aqueous phase was removed and underwent an isopropanol precipitation (incubated for 10 mins (4°), centrifuged for 10 mins at 12,000rcf (4°), and washed with 70% ethanol). The resultant RNA pellet was resuspended in 40µl DEPC H<sub>2</sub>O and underwent DNase 1 treatment to remove any contaminating DNA.

Total RNA concentration was determined by nanodrop spectrometry and 12.5µg RNA from each sample was treated with DNase 1 (1-unit DNase 1: 2.5 µg RNA) for 40 mins at 37°C. The DNase treated RNA underwent a standard ethanol precipitation and the resultant pellet was resuspended in 25µl DEPC H<sub>2</sub>O.

The untreated and DNase1 treated RNA were tested on a 2% agarose gel to confirm that the DNase treatment had worked.

### **3.14 Total RNA sequencing:**

The prepared RNA from untreated and mimosine-treated EJ30 cells underwent RNA sequencing at the Francis Crick Institute. Prior to sequencing, rRNAs were removed. Libraries were generated using the KAPA RNA RiboErase HyperPrep preparation kit according to the manufacturer's instructions and the sequencing was performed using the Illumina HiSeq 4000 platform (75bp SR) according to the manufacturer's instructions. Library preparation and rRNA removal was carried out by R. Jones. T.Krude mapped and aligned the sequencing files to the human hg38 genome.

## **Chapter 4: The development of density-substitution initiation-site sequencing (ds-iniSeq)**

### **4.1 Introduction**

In chapter one, I examined the large body of work that exists on human DNA replication origins, with particular focus on NGS methods available to interrogate the features involved in replication origin specification and activation. While much has been elucidated, the precise characteristics that are necessary and sufficient for the specification of origin sites remains unclear (section 1.3)

I also discussed the development of the NGS sequencing method for origin identification, iniSeq (section 1.5), which utilised an *in vitro* replication reaction from the established human cell-free system (section 1.4). During this 15-minute replication reaction, a digoxigenin-tagged dUTP was introduced into newly synthesised DNA (ie replicated DNA). Following this reaction, total DNA was fractionated and separated via immunoprecipitation, using anti-digoxigenin antibodies. The total genomic input DNA and the separated newly synthesised digoxigenin-tagged DNA were subjected to NGS. Newly synthesised/replicated DNA was enriched at discrete sites, which allowed for the identification of >20,000 human replication origins (1).

Initially, I attempted to convert the iniSeq method (in accordance with Langley *et al* (1) to our lab, with the aim to use it for further analysis of human replication origin specification and activation. Unfortunately, I encountered multiple difficulties with this conversion, including isolating insufficient replicated DNA for NGS. Once I was able to isolate sufficient replicated DNA, our collaborators at the MRC-LMB (G. Guilbaud (GG), J. Sale group) performed NGS. Origin sites were present in the replicated DNA, but they were not sufficiently enriched above the background level of sequencing. Eventually, GG established that the high-fidelity DNA polymerases required for the sequencing protocol were inefficient in overcoming the digoxigenin incorporation in the replicated DNA; this was not a problem encountered during the initial development of iniSeq.

I concluded that the original iniSeq method was not viable for the analysis of DNA replication origin specification and must be modified to enable such analyses. This led to the development of density-substitution initiation-site sequencing (ds-iniSeq) method, based upon the principle of density substitution for the separation of replicated and unreplicated DNA. This new method improves on the original iniSeq method as it allowed me to not only determine the specification, but also the determination of the relative activities of DNA replication origins, thus, adding an additional layer to the analysis of human replication

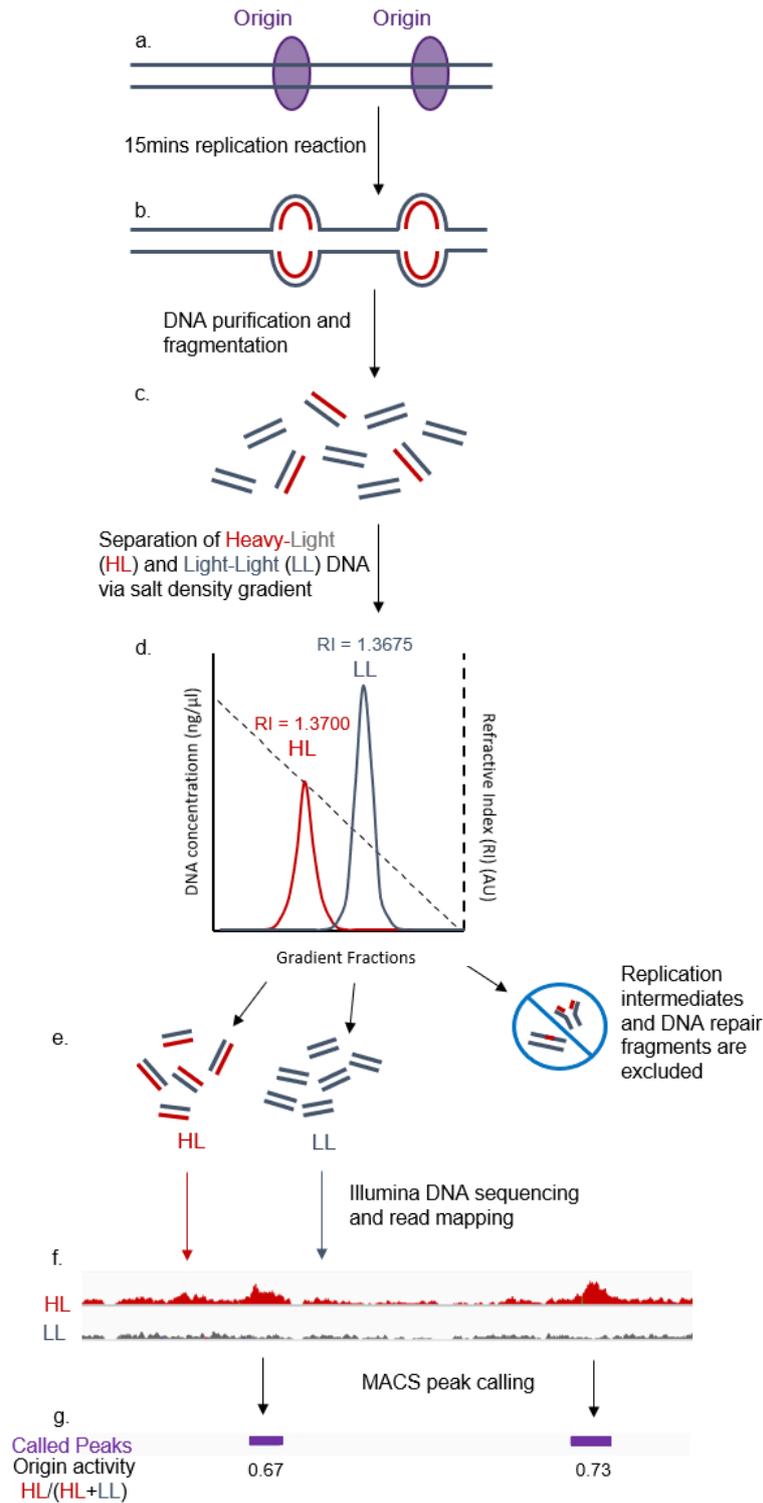
origins. This work was undertaken in collaboration with the J. Sale group (MRC-LMB, Cambridge) and the J. Smith group (Francis Crick Institute, London).

In this chapter, I describe the development of the novel ds-*iniSeq* method and the parameters required to make it a viable and reliable method for replication origin identification and their subsequent analysis.

## **4.2 Results and Discussion**

### *4.2.1 The ds-*iniSeq* method*

The ds-*iniSeq* method (schematic in Fig.4.1) is based on density substitution of semi-conservative DNA replication (established by Meselson and Stahl (2)), whereby a heavy bromo-dUTP (BrdUTP) is incorporated into newly synthesised/replicated DNA instead of lighter dTTP. The replicated heavy DNA can be separated from unreplicated light DNA via density equilibrium centrifugation and then sequenced.



**Figure 4.1:** A schematic of ds-iniSeq method. The potential replication origin sites (a) undertook DNA replication initiation and limited elongation in a 15 minute replication reaction (b), under standard conditions (synchronised G1 phase nuclei were incubated with cytosol from proliferating cells and a mixture of dNTPs, NTPs and an ATP regenerator). The newly synthesised DNA incorporated the heavy BrdU (b). Total genomic DNA was fragmented to produce a mixture of replicated heavy-light (HL) and unreplicated light-light (LL) DNA (c), which was separated on a  $\text{Cs}_2\text{SO}_4$  density gradient (d). Purified HL and LL DNA, which was separated from replication intermediates, DNA repair fragments (e) and RNA (RI = 1.3814 AU), underwent Illumina sequencing and mapping to the human hg38 genome (f). The resultant files underwent MACS peak calling and read count quantitation, to identify origin sites and their corresponding relative origin activities

In this method, DNA replication origins on the human genome (Fig.4.1a) fired during a short 15-minute *in vitro* replication reaction, whereby late G1-phase template nuclei were incubated (37°C) with cytosol from asynchronous HeLa cells (human cytosol), a selection of NTPs, dNTPs and an ATP-regeneration system (phosphocreatine/creatine kinase), in the presence of a physiological buffer. The nuclei overcame the effect of the compound used for synchronisation (mimosine) and DNA replication synthesis began. In this reaction, the light dTTP was entirely replaced with the heavy nucleotide, BrdUTP, which was incorporated into the newly synthesised DNA (Fig.4.1b).

The replication reaction was promptly stopped, and the DNA was purified, using a proteinase K treatment. Total genomic DNA was sonicated resulting in 100-1000bp DNA fragments (Fig.4.1c). At this point the newly synthesised DNA consisted of one heavy and one light DNA strand (HL DNA), whereas the unreplicated DNA consisted of two light DNA strands (LL DNA).

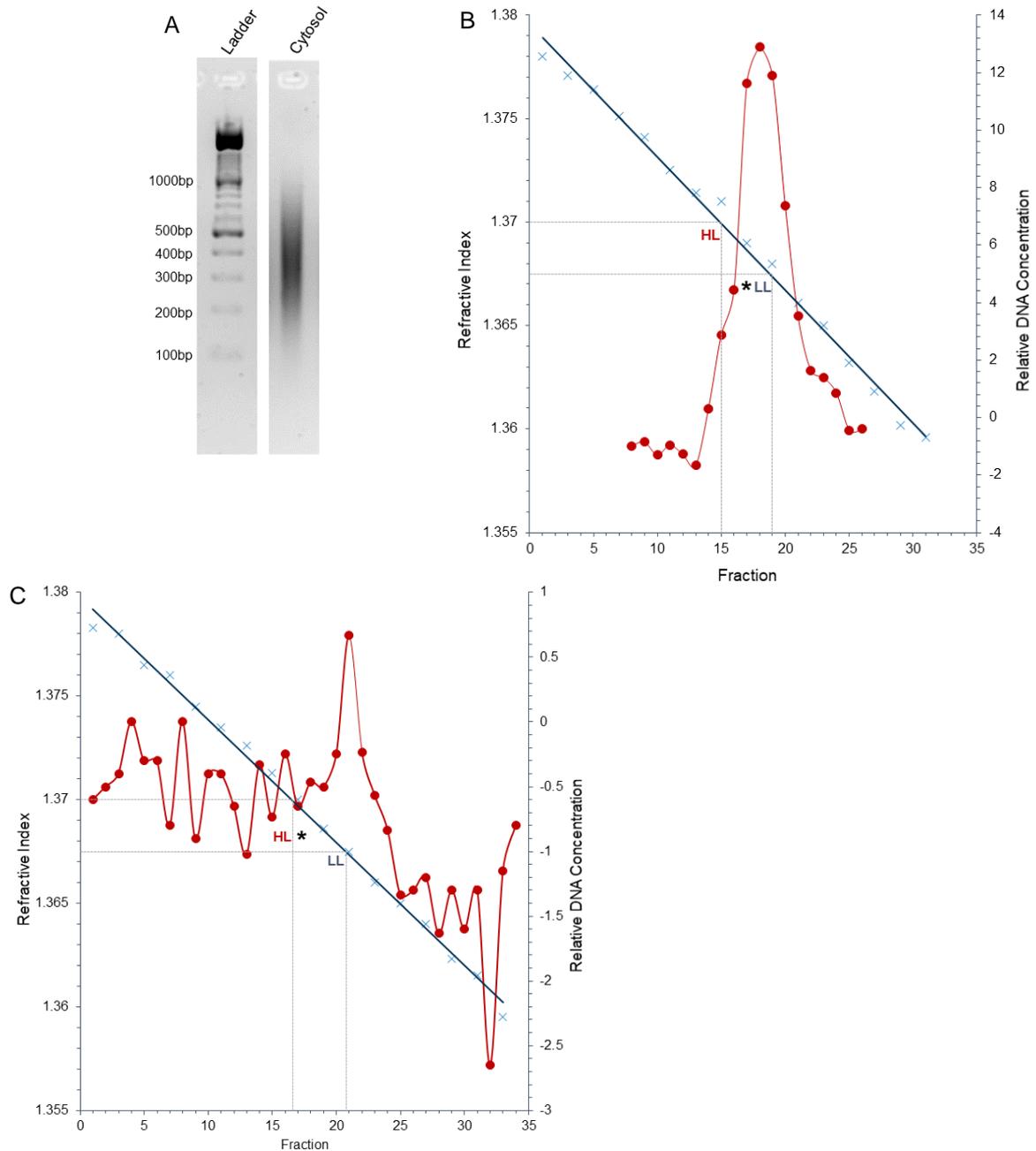
The replicated HL and unreplicated LL DNA were separated via density equilibrium centrifugation using a caesium sulphate gradient (Fig.4.1d). HL DNA was present in fractions at salt concentrations with a refractive index (RI) of 1.3700 and the LL DNA was present in fractions at salt concentrations with a RI of 1.3675. The RNA was separated away from the DNA in a fraction at a RI of around 1.3814 (3). Advantageously, replication intermediates and DNA fragments that underwent damage repair were excluded from both the isolated HL and LL DNA as they have intermediate RIs.

Unreplicated LL DNA was isolated from the first density gradient. The HL DNA fractions from the first gradient were isolated and subjected to a second density gradient for further separation from contaminating LL DNA. The replicated HL DNA was isolated from the second gradient. This resulted in purified and separated HL and LL DNA pools (Fig.4.1e). These pools were desalted and underwent Illumina NGS.

The sequencing data were mapped and aligned to the human hg38 genome (Fig4.3f) and subjected to Model-based Analysis of ChIP-Seq (MACS) peak calling to identify sites and regions that had enrichment of replicated HL DNA over unreplicated LL DNA. The relative amount of sequenced HL and LL DNA was quantified at the enriched sites/regions called by MACS (normalised by per million reads to ensure comparability between experiments), using the SeqMonk software (Babraham Institute (Cambridge)(4)). With these values, the ratio of replicated DNA to total DNA (HL/(HL+LL)) was calculated for each called site/region in order to attribute a relative origin activity value to each called origin (Fig4.3g)

#### *4.2.2 Preliminary ds-*ini*Seq experiment*

Using the new ds-*ini*Seq method (Fig.4.1) I carried out a preliminary run, together with T. Krude (TK)(Department of Zoology, Cambridge) and GG (MRC-LMB, Cambridge), to establish proof of concept. I performed a standard 15-minute replication reaction and purified and fragmented total genomic DNA. A sample of the fractionated DNA was subjected to gel electrophoresis (2% agarose gel). The replicated HL and unreplicated LL DNA were separated on two salt density gradients (performed by TK and GG at the MRC-LMB) (Fig.4.2).



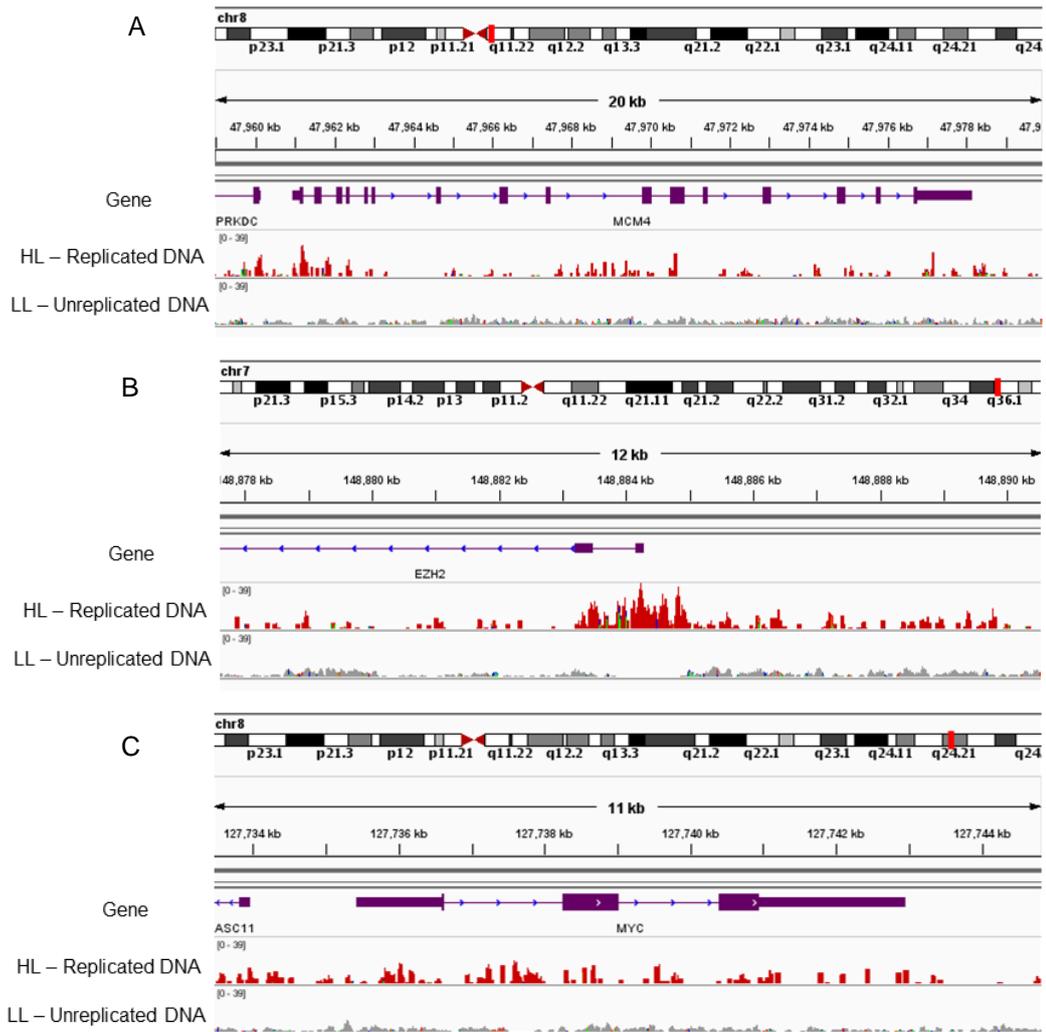
**Figure 4.2:** The 2% agarose gel (A) shows the successful fragmentation of total genomic DNA (500ng loaded) from the preliminary ds-*iniSeq* experiment (“cytosol”). A standard molecular weight ladder (2 $\mu$ l) was also loaded. The first (B) and second (C – HL fractions from the 1<sup>st</sup> gradient were loaded onto the 2<sup>nd</sup> gradient); density gradients from the preliminary ds-*iniSeq* experiment show salinity across the gradient fractions, as determined by refractometry (refraction index values (blue) for each fraction is indicated on the left vertical axis), and the relative DNA concentration (red) across the gradient fractions, as determined by nanodrop spectrophotometry is indicated on the right vertical axis. For both density gradients (B&C) the refractive indices for HL and LL are indicated with grey dotted lines and the appropriate labels. The \* indicates the fractions from which samples were selected, desalted, and sent for Illumina sequencing.

Figure 4.2A shows that the total DNA was successfully fragmented to sizes between 100 and 1000bps, with a peak around 300bps.

The RI of the fractions obtained from the first density gradient (Fig.4.2B) showed a reasonable separation of different salt concentrations (RI) across the fractions, indicating the formation of a linear gradient during equilibrium centrifugation. The relative DNA concentrations (determined by nanodrop spectrophotometry) (Fig.4.2B) showed a large LL DNA peak (RI 1.3675), that overlapped with a smaller HL DNA peak (RI 1.3700). Therefore, this HL peak appeared as a small “shoulder” of the LL peak. The LL DNA was selected from fractions on the right-hand side of the LL peak, which were pooled, desalted, and underwent Illumina sequencing. Fractions of the HL DNA “shoulder” were pooled and run on a second density gradient, to remove contaminating LL DNA.

The second density gradient (Fig.4.2C) also showed a reasonable separation of fractions at different salt concentrations, indicating the formation of a linear gradient during equilibrium centrifugation. The relative DNA concentrations showed a LL DNA peak and only a small relative DNA concentration at the RI for HL DNA. The fractions containing HL DNA were pooled, desalted, and underwent Illumina sequencing. The Illumina sequencing was carried out by our collaborators at the MRC-LMB (Cambridge) (library preparation and sequencing performed by GG (J. Sale group)). Overall, I was able to successfully separate HL and LL DNA.

The sequenced replicated HL and unreplicated LL DNA were mapped and aligned to the human genome (hg38), by TK. The resultant bam files were imported into the integrated genome viewer (IGV) and I examined the HL and LL DNA distributions at the established MCM4, MYC and EZH2 origin (Fig.4.3).



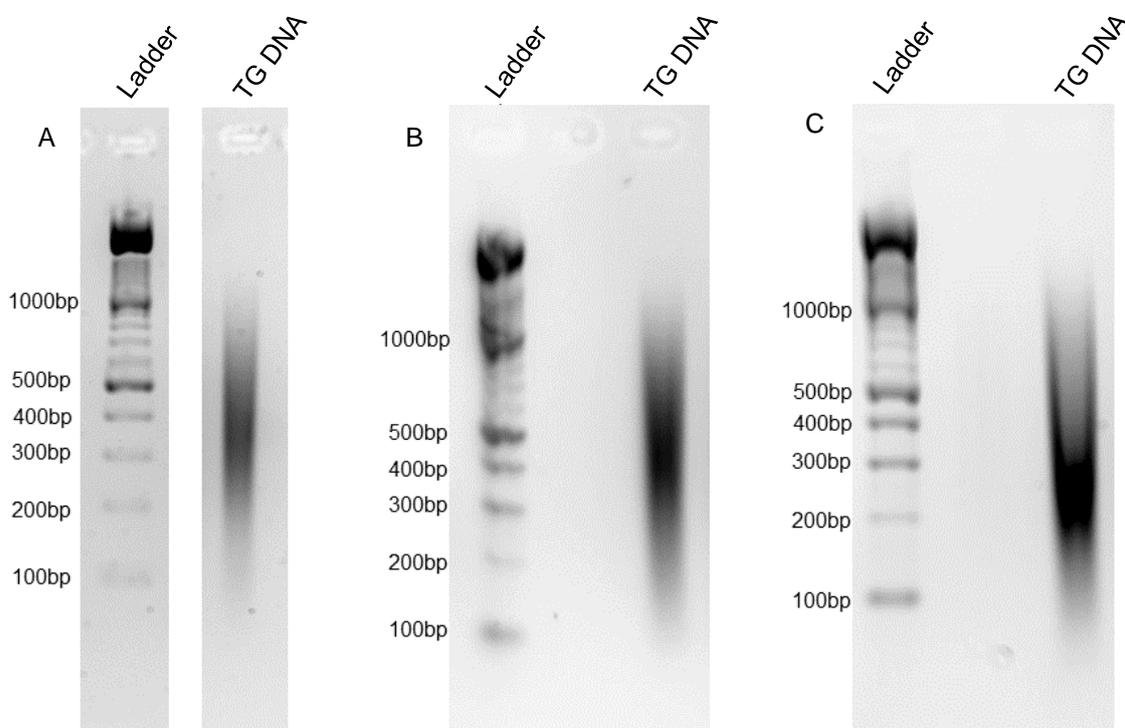
**Figure 4.3:** The IGV images of the mapped sequencing data generated from the preliminary ds-*iniSeq* experiment, where the replicated HL DNA is indicated in red, the unreplicated LL DNA is indicated in grey and the reference genes are indicated in purple. The chromosome and position (bright red marker on the chromosome) on the chromosome are also indicated above the HL and LL profiles. (A) shows the above for the well-established origin at the MCM4 gene promoter (“MCM4”), (B) shows the highly activity replication origin present at the EZH2 gene promoter (“EZH2”), and (C) shows the well-known replication region/zone present at the MYC gene (“MYC”).

The origin sites at the MCM4 and EZH2 promoters (Fig.4.3A&B) showed a relative accumulation of reads in HL DNA and a relative partial depletion of reads in LL DNA (with a complete depletion at the highest HL read accumulation regions). The replication initiation zone at the MYC gene (Fig.4.3C) showed a relative accumulation of reads in HL DNA and low read accumulation in LL DNA with areas of relative partial/moderate depletion of read accumulation. However, this depletion was also seen in other areas that possessed no HL DNA enrichment. The relative depletion of reads in LL DNA may have resulted from low sequencing read depth coverage.

The preliminary data (Fig.4.2/3) demonstrated that the ds-iniSeq method successfully detected human DNA replication origins. I performed three full replicates of the standard ds-iniSeq reactions to determine the parameters for origin peak calling and assess the reproducibility of the method to identify and characterise activated origins.

#### 4.2.3 Production of ds-iniSeq replicates

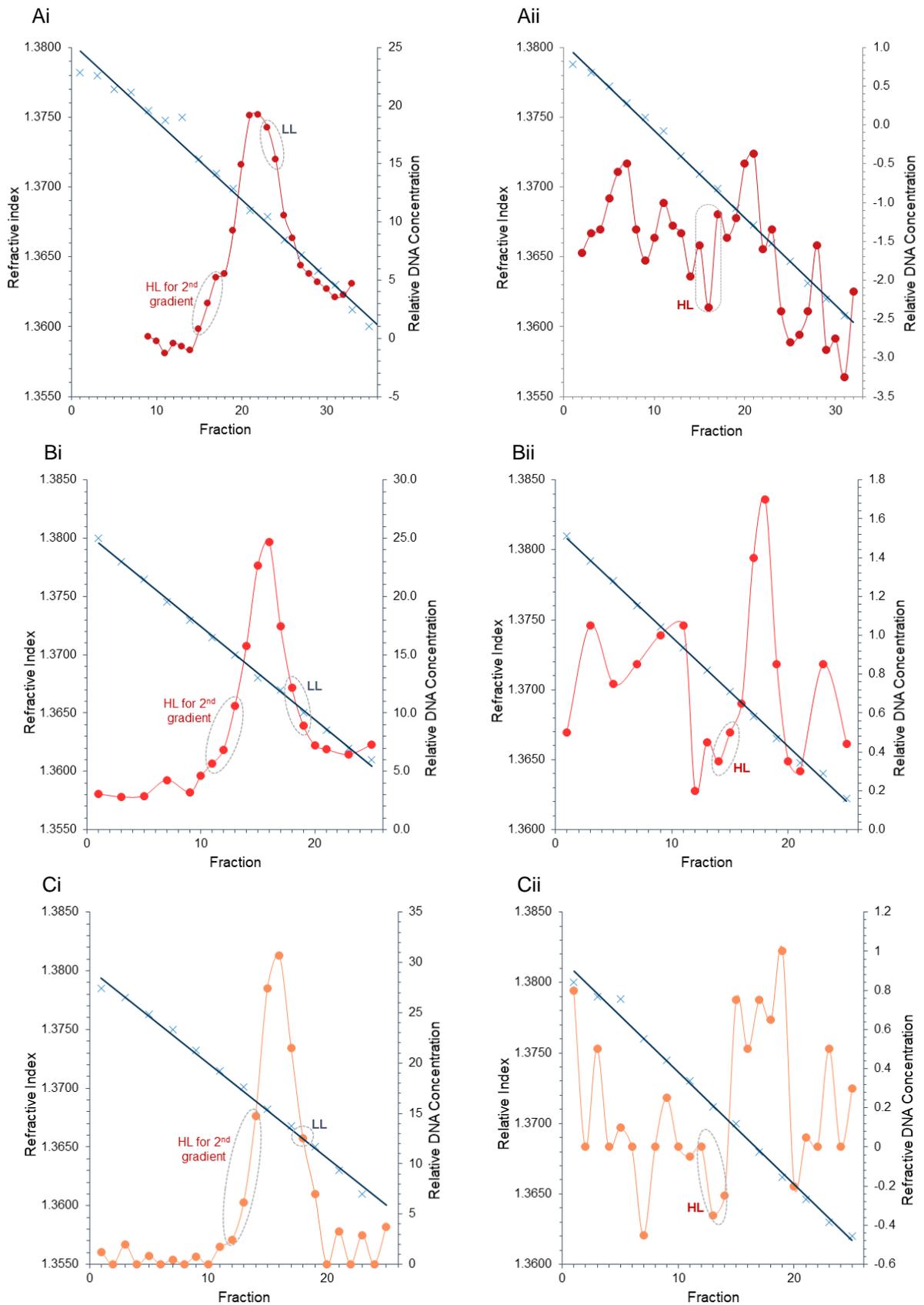
I performed the same standard DNA replication reaction and purified and fractionated total genomic DNA, in three separate experiments. DNA samples for each ds-iniSeq replicate were subjected to gel electrophoresis (2% agarose gel) (Fig.4.4).



**Figure 4.4:** The 2% agarose gels that were used to assess the quality of fragmentation of the sonicated total genomic DNA (“TG DNA”) generated from the standard ds-iniSeq reaction, of replicates 1 (A), 2 (B) and 3 (C). The molecular weight ladders were also loaded on the gel and 100 – 1000bp were indicated (“Ladder”). 500ng of each sample for each replicate were loaded on each gel.

The agarose gels for replicates 1 (Fig.4.4A), 2 (Fig.4.4B) and 3 (Fig.4.4C) showed that the total genomic DNA was successfully fragmented to 100 to 1000bps in length, with peaks at ~350bps (replicate 1), ~400bps (replicate 2) and ~200bps (replicate 3). These marginal differences were unlikely to make a substantial difference to subsequent steps.

I then separated the HL and LL DNA on density gradients (Fig.4.5).



**Figure 4.5:** The first (i) and second (ii) density gradients for replicates 1 (A), 2 (B) and 3 (C) showed salinity across the gradient fractions, as determined by refractometry (refraction index values (blue) for each fraction is indicated on the left vertical axis), and the relative DNA concentration (dark red for replicate 1, cherry red for replicate 2, and orange-red for replicate 3) across the gradient fractions, as determined by nanodrop spectrophotometry is indicated on the right vertical axis. For the 1<sup>st</sup> gradients, the LL fraction that were used for sequencing and the HL fractions that were used for the 2<sup>nd</sup> gradient are indicated. For the 2<sup>nd</sup> gradients, the HL fractions that were used to sequencing are also indicated.

The density gradients from replicate 1 (Fig.4.5A) confirmed that the first (Fig.4.5Ai) and the second (Fig.4.5Aii) equilibrium centrifugations provided linear gradients of caesium sulphate salt concentrations (RI) across the fractions. The relative DNA concentrations from the first gradient showed a large LL DNA peak (RI 1.3675) with a smaller “shoulder” at the RI corresponding to HL DNA (1.3700). The unreplicated LL DNA was collected from fractions to the right of the main LL (RI  $\leq$  1.3675) peak (fractions 23&24), desalted and sequenced. The replicated HL DNA was collected from fractions to the left (away from the LL peak) of the HL peak/“shoulder” (RI  $\geq$  1.3700; fractions 15-17).

The pooled HL DNA fractions were run on a second density gradient for each replicate. The second gradient of replicate 1 showed a similar pattern to that of the preliminary experiment. The separation of fractions based on salt concentration (RI) was adequate, indicating a linear gradient formation during equilibrium centrifugation. The relative DNA concentration showed a moderate LL DNA peak (RI 1.3675) and a small HL DNA peak/“shoulder” (RI 1.3700). The HL DNA was collected from fractions to the left of the HL peak/“shoulder” (RI  $\geq$  1.3700; fractions 15-17), desalted and sequenced.

There was a downward trend towards the baseline relative DNA concentration across the course of the gradient (fractions 1-32), which was explained by the decreasing salt concentrations of the fractions when compared to that of the solution used to calibrate the nanodrop spectrophotometer. In the second gradient, the relative concentration measurements of DNA fluctuated more than the first, due to the lower amounts of DNA in the second gradient leading to lower nanodrop sensitivity.

Both gradients were also generated for replicates 2 (Fig.4.5B) and 3 (Fig.4.5C) and showed the formation of a linear gradient of salt concentrations during equilibrium centrifugation. The distributions of relative DNA concentration across the first gradients for both replicates 2 (Fig.4.5Bi) and 3 (Fig.4.5Ci) follow a similar pattern to replicate 1. They underwent the same process of LL (fractions used for sequencing indicated on figure) and HL DNA (fractions used for 2<sup>nd</sup> gradient indicated on figure) selection as replicate 1.

The relative DNA concentrations of the second gradients for replicates 2 (Fig.4.5Bii) and 3 (Fig.4.5Cii) showed greater fluctuations than their respective first gradients. They retained a contaminating LL peak, and HL DNA fractions were selected (indicated on figure) and underwent the same process as replicate 1. All the second gradients possessed a LL DNA peak, highlighting the need to run a second gradient to further remove contaminating LL DNA from the HL DNA. In all cases, I selected HL DNA fractions to the left of the RI for HL DNA for sequencing, to further minimise LL DNA contamination.

However, there is a potential for LL contamination within the HL samples, especially considering the small amount of replicated HL DNA produced when compared to the total LL

DNA within the human cell-free system, upon which ds-iniSeq is based. A rudimentary way in which to assess the degree of LL contamination within the HL sample, is to visualise the sequenced data at sites that are consistently established as non-replicating zones/regions. Although, the degrees of LL contamination may not be consistent across the whole genome and this would not provide an objective measure.

One potential method to further minimise the contamination is to run further caesium sulphate gradients with the pooled fractions at the refractive index for HL DNA. This may only reduce the degree of LL contamination rather than remove it entirely. In the subsequent data analyses, the LL DNA that was collected from the first gradient was used as a comparison to the HL DNA in an attempt to account for LL contamination/background within the HL samples. Yet, these LL DNA samples may not fully account for the LL DNA potentially contaminating the HL samples.

The DNA in these ds-iniSeq experiments were fragmented to 100-1000bp fragments. These fragments may differ in composition. The GC content of the human genome, in 100bp fragments, ranges from 35% to 60% (5). It is reasonable to assume that there is a similar range of GC content within the fractionated samples here. It has previously been shown that GC content of DNA impacts its buoyant density and, consequently, at which refractive index it is found. From calibrations using Caesium Chloride density gradients, it was established that the higher the GC content of DNA, the higher its buoyant density (6,7). As these findings were established on a CsCl gradient, and these ds-iniSeq experiments were conducted using a Cs<sub>2</sub>SO<sub>4</sub> gradient, a similar calibration of GC content in a Cs<sub>2</sub>SO<sub>4</sub> gradient would be a useful addition. However, it remains likely that a similar relationship would be found between GC content and buoyant density. Although whether the refractive index of the highest GC containing fragments of LL DNA is similar to that of HL DNA remains unelucidated. It is possible that any potential contaminating LL DNA in the HL samples is comprised of a higher GC content than the main LL peak.

To address this issue, I would propose that, in addition to establishing the buoyant density of LL DNA with differing GC content on a Cs<sub>2</sub>SO<sub>4</sub> gradient, an additional ds-iniSeq experiment should be run in the absence of Br-dUTP to allow for the assessment of the amounts of LL DNA in each fraction of both the first and second gradients. Subsequent sequencing of each DNA containing fraction would then enable the elucidation of the GC content across the gradient. This would establish to what extent the LL DNA may contaminate the HL samples and if there is a bias for a higher GC content than the LL DNA samples account for in subsequent assessments.

These replicates and the preliminary sample established that this method of separation of replicated HL and unreplicated LL DNA was highly reproducible, reliable and viable for further experimentation.

I isolated replicated HL and unreplicated LL DNA for all three replicates and desalted, precipitated, and resuspended them in water. The final DNA concentrations and amounts of the samples, determined by nanodrop spectrophotometry, are shown in table 4.1.

Sample	Concentration (ng/ $\mu$ l)	Volume ( $\mu$ l)	Amount (ng)
<b>Replicate 1 HL</b>	4.8	31	148.8
Replicate 1 LL	92.1	32	2,947.2
<b>Replicate 2 HL</b>	19.7	20	394.0
Replicate 2 LL	63.6	20	1,272.0
<b>Replicate 3 HL</b>	23.3	20	466.0
Replicate 3 LL	88.0	20	1,760.0

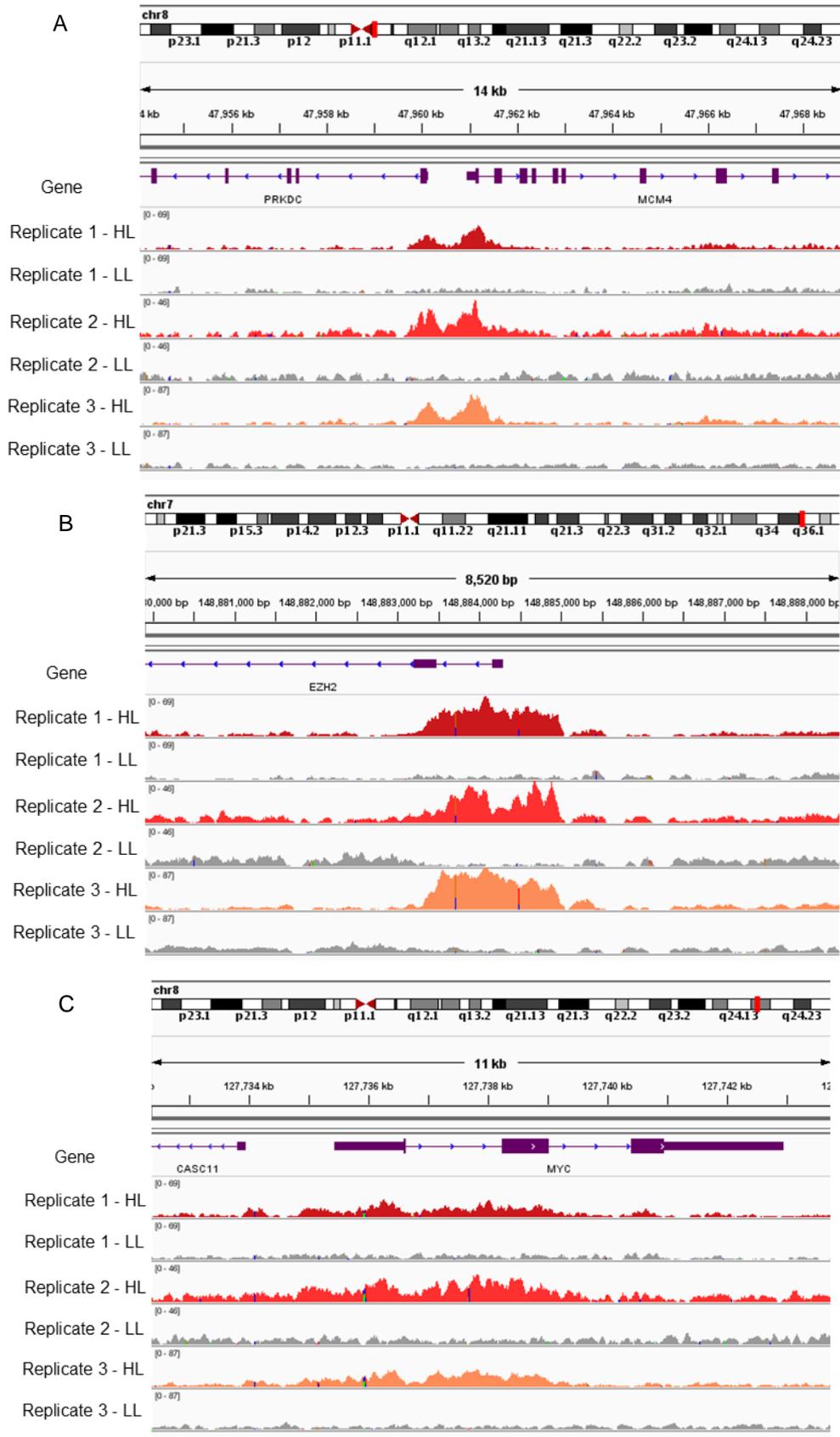
**Table 4.1:** The final concentrations, volumes and amounts of both HL and LL DNA obtained from the replicates 1, 2 and 3. The HL and LL fractions highlighted for sequencing in figure 4.7 were pooled (where needed), desalted and underwent an ethanol precipitation and resuspended in DEPC H<sub>2</sub>O. The concentrations of DNA were determined through nanodrop spectrophotometry.

There were substantial amounts of DNA in both replicated HL and unreplicated LL DNA for all three replicates with consistently higher amounts of unreplicated LL DNA (~1,300-3,000ng), compared to the replicated HL DNA (~150-500ng) (Table 4.1). All three replicates contained sufficient HL and LL DNA to successfully undergo library preparation and Illumina sequencing. This demonstrates that the ds-iniSeq method overcame the issue of insufficient DNA that arose with the original iniSeq method.

#### 4.2.4 NGS sequencing

The HL and LL DNA samples underwent Illumina sequencing by our collaborator's NGS facility at Francis Crick Institute (London) (library preparation by K. Dingwell (replicate 1 HL) and R. Jones (replicate 1 LL, replicates 2 HL&LL and 3 HL&LL), J. Smith group). All duplicates and non-unique reads were removed and final .bam files were generated. The files were mapped and aligned to the human hg38 genome by TK.

I imported these files into the integrated human genome viewer (IGV), to visualise the data at the known origin sites/regions at the MCM4 and EZH2 promoters and the MYC gene (Fig.4.6).



**Figure 4.6:** The IGV images show the mapped sequencing data generated from the replicated HL DNA (dark red for replicate 1, cherry red for replicate 2, and orange-red for replicate 3), the unreplicated LL DNA (grey for all replicates) and the reference genes (purple). The chromosome and position (bright red marker on the chromosome) on the chromosome are indicated above the HL and LL profiles. (A) The “MCM4” origin. (B) The highly active “EZH2” replication origin. (C) The “MYC” replication initiation zone.

These well-established and prominent DNA replication origins consistently possessed an accumulation of reads at these sites/regions in the replicated HL DNA but had not in the unreplicated LL DNA (Fig.4.6).

Replicates 1, 2 and 3 all showed a defined double peak, with one peak overlapping the PRKDC gene and the other overlapping the MCM4 promoter in the replicated HL DNA whilst no peak was present there in the unreplicated LL DNA (Fig.4.6A). The origin site found at the EZH2 promoter (Fig.4.6B) was consistently observed for all three replicates, as replicated HL DNA was highly enriched at this site whilst the unreplicated LL DNA was not enriched. Notably for replicates 2 and 3, there was a reduction in the read accumulation of unreplicated LL in the region corresponding with the EZH2 origin, compared to the background of the adjacent neighbouring regions. This may indicate a partial depletion of unreplicated DNA in favour of replicated DNA. Finally, the wider MYC replicating initiation zone (Fig.4.6C) also showed read accumulation in the replicated HL DNA with no enrichment in the corresponding region of the unreplicated LL DNA, for all three replicates. One can also observe the presence of reads adjacent to the established origin peaks in the HL DNA samples. These may indicate the presence of contaminating LL DNA within the HL samples.

These three example origins demonstrate the ability of the ds-iniSeq method to detect DNA replication at known replication origins and show concordance between all three replicates, again suggesting reproducibility between replicates with the ds-iniSeq method.

#### *4.2.5 Origin calling by MACS*

Figures 4.2–4.6 demonstrated that the ds-iniSeq method reliably and reproducibly generated samples that showed an enrichment of read accumulation at the sites of known replication origins/regions, that was specific for replicated HL DNA. I then established a method to accurately call replication origin sites/regions genome-wide in all three replicates.

The MACS peak caller is traditionally used for ChIP-Seq analyses (8). It is based on the principle of comparing regions of the genome that possess enriched read accumulation in a ChIP sample to an unaltered input sample. The ds-iniSeq method utilises a comparable methodology, where the “traditional” ChIP sample is replaced by replicated HL DNA (with specific enrichments of read accumulation at replication origins) and the “traditional” input sample is replaced by unreplicated LL DNA (with negligible/no enrichment or depletion in read accumulation at the corresponding sites). In essence, ds-iniSeq generates datasets equivalent to ChIP-seq and, as such, the MACS peak caller could be applied to calling replication origin sites/regions.

I used the Babraham Institute’s bioinformatic software SeqMonk’s MACS function and quantification tool (4) to determine and optimise the parameters for peak calling and

ultimately generate the finalised lists of origins for these replicates produced here and any subsequent ds-*ini*Seq experiments. For all the MACS peak calling analyses performed, the mapped and aligned HL files were used as the “ChIP” and the mapped and aligned LL files were used as the “Input”.

The parameters I determined/optimised are as follows:

1. The effect of read numbers of the replicated HL and unreplicated LL samples on the number of origins called by MACS peak caller.
2. Which p cut-off value should be used for MACS peak calling? – MACS determines peaks by establishing where the replicated HL DNA is significantly enriched above the unreplicated LL DNA. The p cut-off value is the value at which this difference becomes significant.
3. At what distance should neighbouring MACS called sites be grouped and classed as one origin? – There are known replication zones/regions (9,10) which may have been identified as multiple discrete sites located close together by MACS, as MACS is best suited to discrete sites (8). It is possible that two neighbouring MACS called sites should have been considered as one larger origin.

*1. Origin calling – does the HL & LL read number affect the number of called origins?*

In order to determine to what extent, if any, read number of the replicated HL and unreplicated LL samples effected the number of origins called by MACS peak calling, I performed a series of read number titrations, whereby I measured the effect of different numbers of reads on MACS peak calling (Fig.4.7).

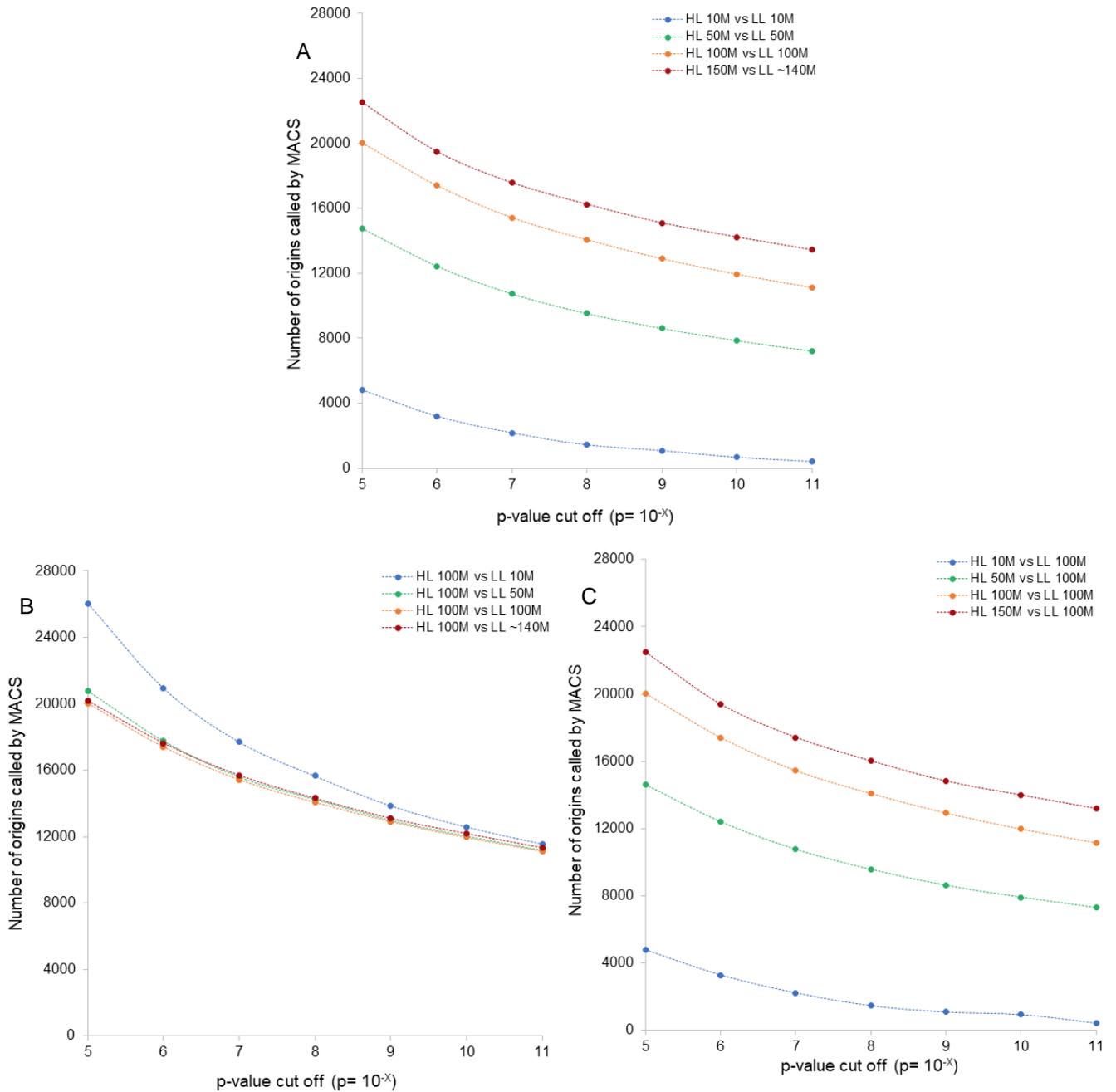
From the original sequencing files, TK generated .bam files for the HL and LL samples with defined read numbers; 10 million reads (10M), 50 million reads (50M), 100 million reads (100M) and 150 million reads (150M). In the case of the replicate 1 LL sample, the original bam file contained only 140,495,985 reads (~140M) and had to be used instead of a 150M file.

I performed a “matched-pairs” titration, where I used MACS peak caller at different p cut-off values to call replication origin sites in the following conditions; HL10M vs LL10M, HL50M vs LL50M, HL100M vs LL100M and HL150M vs LL~140M (Fig.4.7A).

I performed a “fixed HL” titration, where I used MACS peak caller at different p cut-off values to call replication origin sites when the read number of HL files was fixed at 100M. The read number of LL DNA varied, and the conditions were as follows: HL100M vs LL10M, HL100M vs LL50M, HL100M vs LL100M and HL100M vs LL~140M (Fig.4.7B).

I performed a “fixed LL” titration, where I used MACS peak caller at different p cut-off values to call replication origin sites when the read number of LL files was fixed at 100M. The read

number of HL DNA varied, and the conditions were as follows: HL10M vs LL100M, HL50M vs 100M, HL100M vs LL100M and HL150M vs LL100M (Fig.4.7C).



**Figure 4.7:** The sequenced files generated for the replicated HL and unreplicated LL of replicate 1 were of differing read numbers. Files with defined numbers of reads were generated for both the HL and LL and several titrations were performed. Origins were called using MACS peak caller at increasingly stringent p cut-off values (determined at which point the HL was significantly enriched above LL), using HL and LL data with differing total read numbers. (A) shows the “matched-pairs” titrations where the number of reads in HL = that of LL. The following conditions are shown: HL 10M vs LL 10M (blue), HL 50M vs LL 50M (green), HL 100M vs LL 100M (orange) and HL 150M vs ~140M (red). (B) shows the “fixed HL” titrations where the number of reads in HL = that of LL. The following conditions are shown: HL 100M vs LL 10M (blue), HL 100M vs LL 50M (green), HL 100M vs LL 100M (orange) and HL 100M vs ~140M (red). (C) shows the “fixed LL” titrations where the number of reads in HL = that of LL. The following conditions are shown: HL 10M vs LL 100M (blue), HL 50M vs LL 100M (green), HL 100M vs LL 100M (orange) and HL 150M vs 100M (red).

The “matched-pairs” titration (Fig.4.7A) showed that the number of origins called by MACS peak caller increased as the number of reads increased at all p cut-off values. The 10M condition generated the fewest origins, with increasing numbers of origins called as the read number increased. As the p cut-off value became smaller (ie more stringent), the number of called origins reduced, but the pattern of number of called origins increasing with read number remained at each p cut-off value.

The number of reads of the HL and LL files clearly impacted the number of origins called by MACS. To distinguish whether the effect of read number on called origin number resulted from the read number of the HL or LL files or both, I performed the “fixed-LL” and the “fixed-HL” titrations.

The “fixed-HL” titration (Fig.4.7B) determined whether the read number of the LL files impacted the number of origins called by MACS. The number of origins called by MACS for HL100M vs LL10M was greater than that of LL50M, LL100M and LL~140M. However, the number of origins called by MACS for HL100M vs LL50M, LL100M and LL~140M were almost identical at every p cut-off value. For all conditions, as the p cut-off value became more stringent, the number of called origins reduced. I concluded that the number of reads in the LL file had a negligible impact on the number of origins called by MACS, above 50M.

The “fixed-LL” titration (Fig.4.7C) determined whether the read number of the HL files impacted the number of origins called by MACS peak caller. The distribution of origin numbers followed exactly the same pattern as the “matched-pairs” titration; as the HL read number increased (whilst the LL read number remained at 100M), the number of called origins increased. Similarly, as the p cut-off value became smaller (ie more stringent), the number of called origins reduced for all conditions.

I concluded that the number of reads present in the HL files did substantially impact the number of origins called by MACS. Thus, when using MACS, the highest number of reads for the HL files should be used, whereas the number of reads for the LL files can vary as long as it remains above 50M. Under these conditions, the same analyses were carried out for replicates 2 and 3 (appendix A4; Fig.A4.1&A4.2) and the same patterns were observed, although the total number of called origins differed.

## *2. Origin calling – which p cut-off value should be used for MACS peak calling?*

As the read number titration demonstrated, the p cut-off value impacted the number of origins called by MACS. I performed a comparative p cut-off value titration for replicates 1 (HL150M vs LL~140M), 2 (HL150M vs LL150M) and 3 (HL150M vs LL150M) (Fig.4.8A).

SeqMonk possesses a read count quantification tool, which enabled me to determine the localised read count accumulation of HL and LL DNA samples at each MACS origin site.

From normalised (per million reads) read count quantification values for the called origin sites in the HL and LL samples, I generated a relative activity value for each origin, through the ratio of the HL read quantification value to the total read quantification value.

Relative origin activity was calculated for each individual called origin as follows:

$(\text{HL read quantification value} / (\text{HL read quantification value} + \text{LL read quantification value}))$

Relative origin activities of 0, 0.5 and 1 indicated complete LL read enrichment over HL, no read enrichment of either HL or LL and complete HL read enrichment over LL respectively.

These activity values remain relative, rather than absolute values for DNA replication activity, as the full extent of replication that had taken place within the system was not accounted for.

Theoretically, one could quantify absolute origin activities by incorporating the extent of replication within the replication experiment. Traditionally, the human cell-free system establishes the degree of DNA replication by determining the percentage of late G1-phase nuclei that undergo replication during an *in vitro* replication reaction. This could be integrated into the relative activities and provide an absolute maximum amount of replication that could have taken place which would, in turn, generate an absolute replication origin activity. For example, should 50% of all nuclei replicate, the maximum absolute origin activity value could be no more than 0.67.

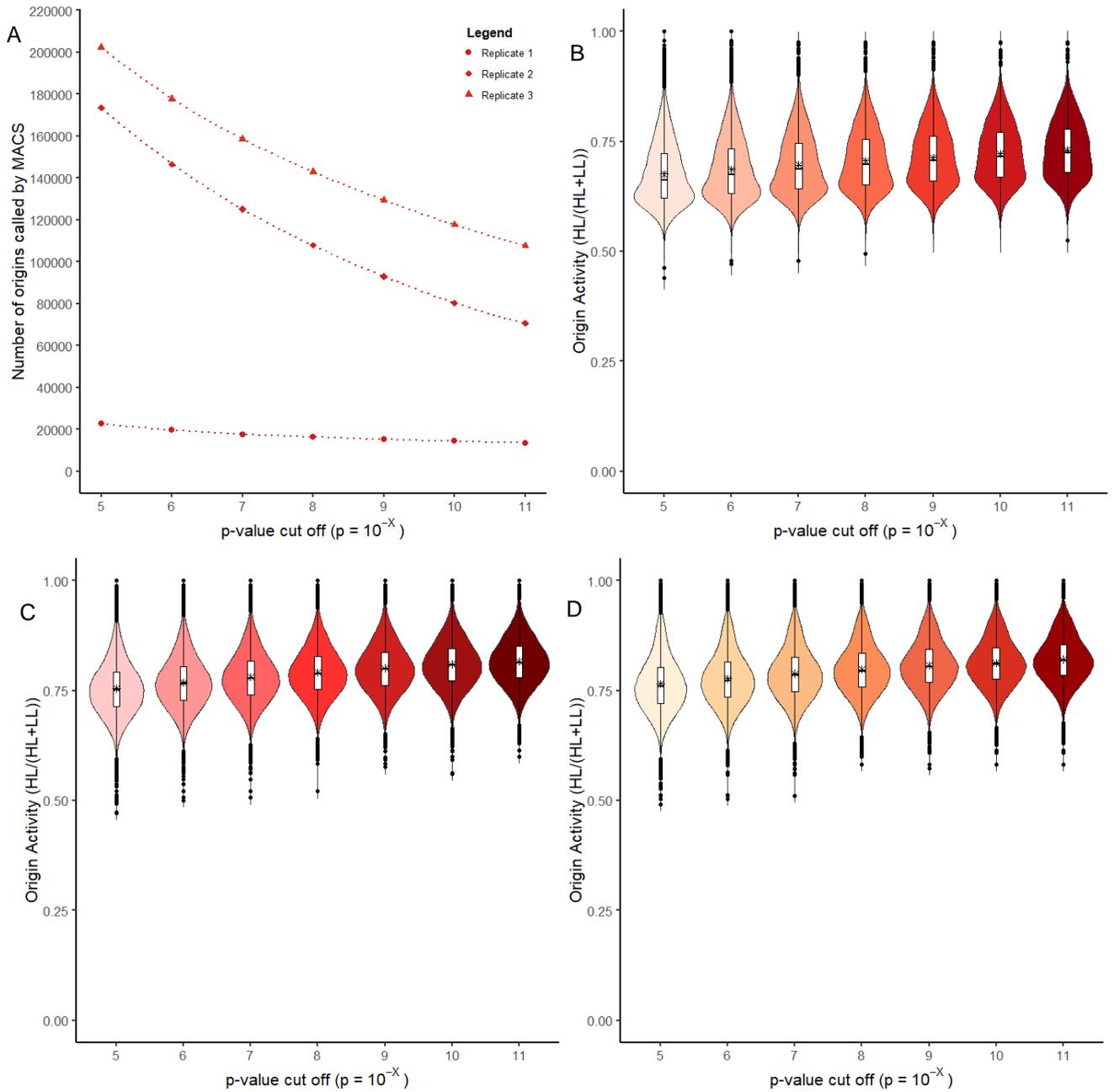
Whereas, with these relative activity values can range from 0 to 1 and as such provides an indication of the probability of an origin firing rather than an absolute quantification of the total DNA replicated. These relative origin activities are valid for usage in all subsequent analyses, but conclusions will not be made based upon the precise numerical values of the origin activities.

Calculating the percentage of nuclei that undergo replication during the ds-iniSeq replication reaction could enable the generation of an absolute quantification of origin activity in these experiments. However, the ds-iniSeq replication reactions are only 15 minutes, which is insufficient time to accurately assess the degree of Br-dUTP incorporation using the traditional confocal microscopy method employed in the human cell-free system. Consequently, one would not be able to obtain an accurate percentage of nuclei undergoing replication in a reaction.

Additionally, the determination of percentage of replicating nuclei using confocal microscopy is dependent upon the subjective assessment of whether a nucleus is replicating or not. Therefore, I would propose assessing the amounts the replicated DNA within the ds-iniSeq experiments themselves. As established earlier, the replicated HL and unreplicated LL DNA are separated using a density substitution gradient and it is possible to quantify the amount of DNA present at the fractions with a RI for HL DNA. This can then be compared to the

known total amount of DNA that underwent fractionation (the input DNA) and produce a resultant percentage of DNA that has replicated within each experiment. This could subsequently be applied to the relative origin activity values and generate an absolute replication origin activity. Unfortunately, I was unable to generate these absolute values and have conducted my data analysis using the relative origin activities.

I generated the relative origin activities of the called origins at each p cut-off value, for replicates 1, 2 and 3 (Fig.4.8B-D).

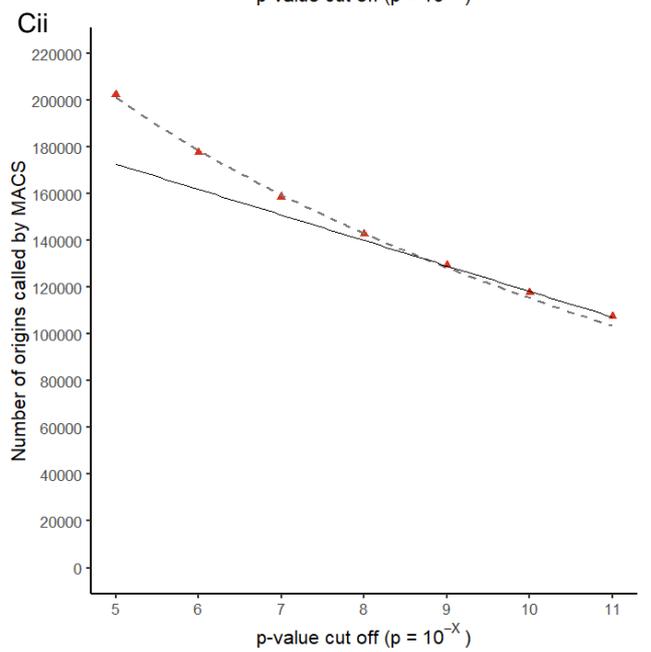
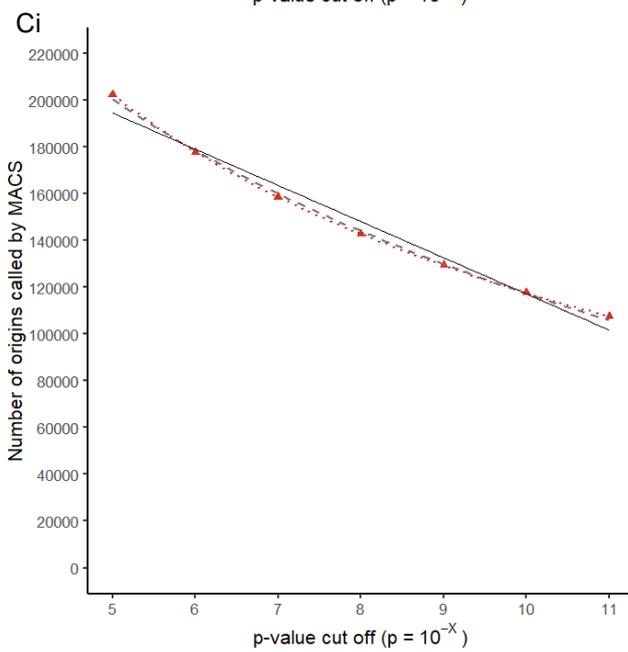
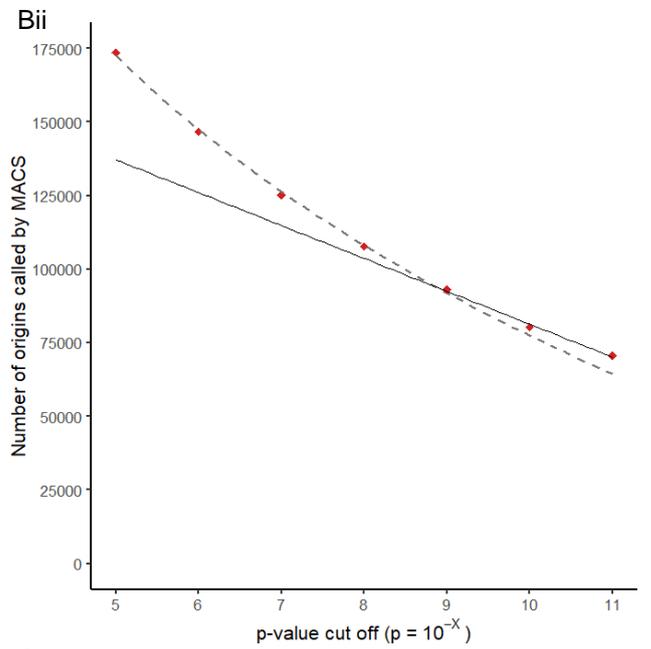
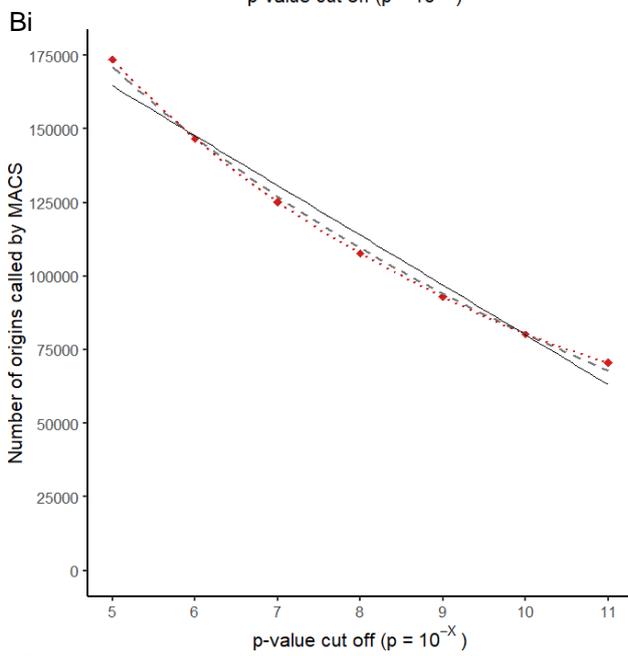
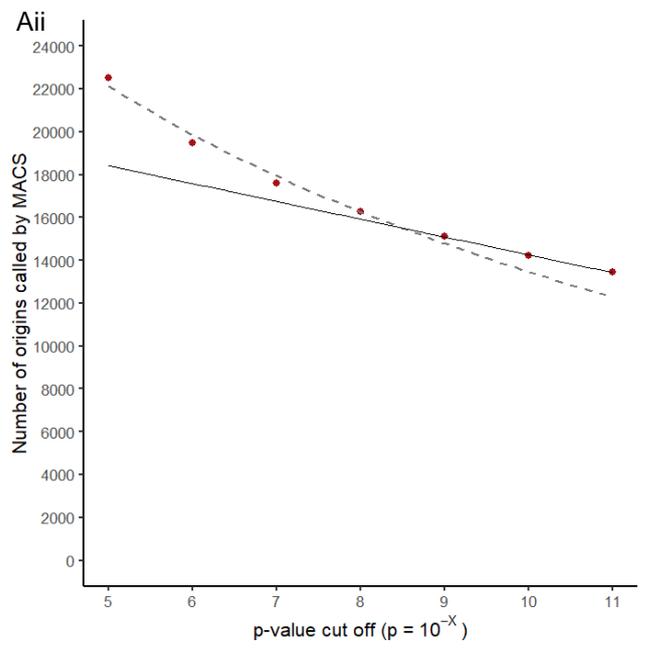
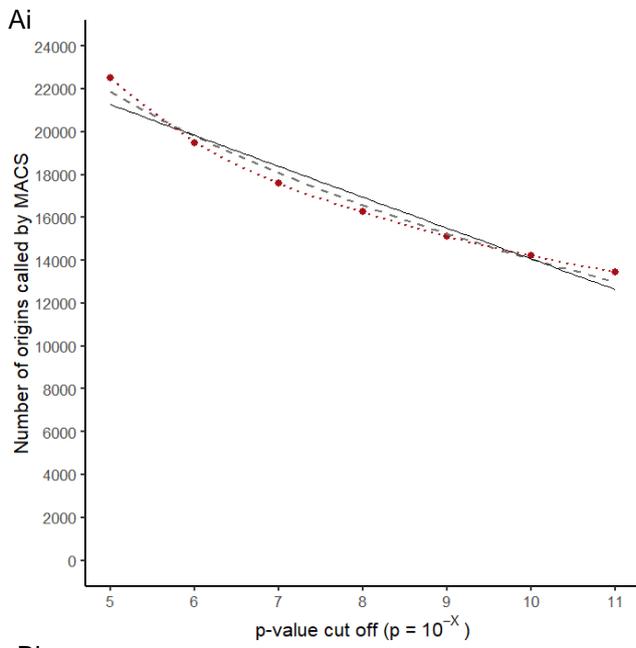


**Figure 4.8:** The number of origins called by MACS peak caller, for replicates 1, 2 and 3, at different p cut-off values (which defined at which point the read accumulations of HL DNA was significantly greater than the corresponding LL DNA), where  $10^{-5}$  was the most lenient/least stringent and  $10^{-11}$  was the most stringent. (A) shows the effect of the increasingly stringent p cut-off values, on the absolute number of origins called for replicates 1 (dark red circle), 2 (cherry red diamond) and 3 (dark orange-red triangle). The origin activities (localised read count accumulation of HL and LL at each origin site which produced a ratio (HL/(HL+LL))) of the origins called at each p cut-off values are plotted for replicates 1 (B), 2 (C) and 3 (D). The interquartile ranges, medians and outliers are indicated as overlying boxplots and the means are indicated with \*.

In all replicates, the number of called origins declined as the p cut-off value decreased ( $10^{-5}$  to  $10^{-11}$ ; ie became more stringent) (Fig.4.8A). This reduction was more substantial for replicates 2 and 3 compared to replicate 1 as were the total number of called origins. The number of origins called in replicate 3 was greater than those called in replicate 2, and the origin number of replicates 2 and 3 exceeded those of replicate 1, at all p cut-off values.

The relative origin activity of origins called in replicates 1 (Fig.4.8B), 2 (Fig.4.8C) and 3 (Fig.4.8D) at different p cut-off values showed that as the selection became more stringent, the relative origin activity generally increased. This was expected, as the more stringent p cut-off values selected for sites with a greater enrichment of HL over LL, and thus a higher relative origin activity.

To determine which p cut-off value to use for origin calling, I examined the relationship between the p cut-off value and the number of called origins for exponentiality and linearity, for each replicate individually (Fig.4.9).



**Figure 4.9:** The numbers of origins called by MACS peak caller at increasingly stringent p cut-off values, where  $10^{-5}$  was the most lenient/least stringent and  $10^{-11}$  was the most stringent, for replicates 1 (A - dark red circle), 2 (B - cherry red diamond) and 3 (C - dark orange-red triangle); for all of the replicates, the read numbers for HL and LL were 150 million reads (replicate 1 LL was ~140 million reads as this was the size of the original file)). For each replicate, the whole data sets (i) are displayed with a trendline joining all data points (red dotted), and fitted to an exponential trendline (in the  $10^{-11}$  to the  $10^{-5}$  direction; a logarithmic trendline in the  $10^{-5}$  to  $10^{-11}$  direction) (grey dashed) and a linear trendline (black solid). For each replicate, part of the data sets (ii) are fitted to an exponential trendline ( $10^{-9}$  to  $10^{-5}$ ; grey dashed) and part are fitted to a linear trendline ( $10^{-9}$  to  $10^{-11}$ ; black solid).

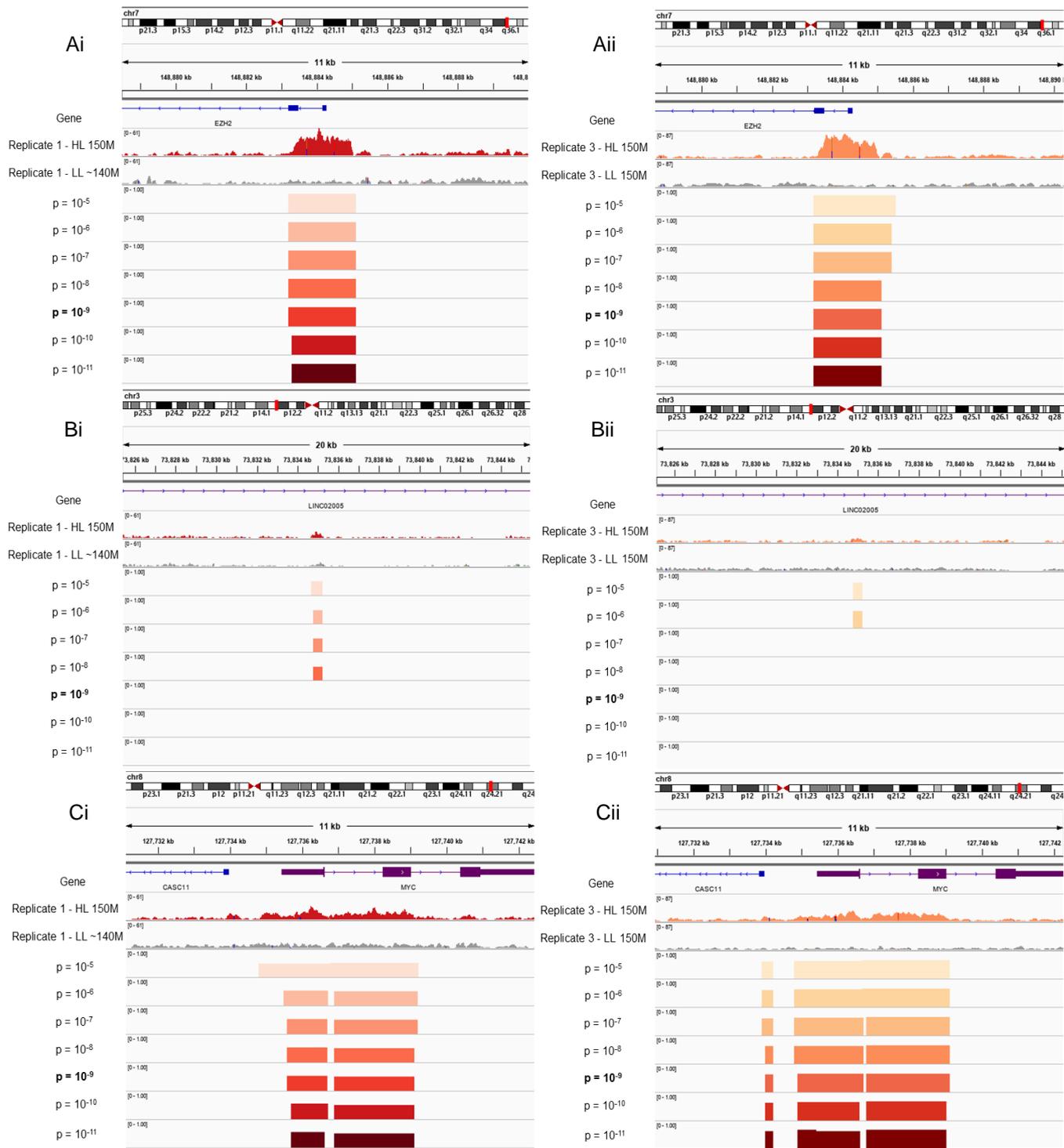
Figure 4.9Ai showed that the relationship between the p cut-off value and number of called origins in replicate 1 was neither exponential nor linear. The (Fig.4.9ii) data points  $p = 10^{-5} - 10^{-9}$  appeared follow an approximately exponentially curve ( $10^{-9}$  to  $10^{-5}$  direction), whereas the curve for the latter points  $p = 10^{-9} - 10^{-11}$  fitted a linear function (for detail see appendix A4, Fig.A4.3).

The exponential relationship at less stringent p cut-off values may indicate the increasing detection of non-specific sites in addition to true origin sites. As a result, it is likely that more false positives were being called as the p cut-off value became less stringent (Fig.4.8B). Whereas the linear relationship indicates the calling of true origins; as MACS became more stringent, origin sites of greater activity (Fig.4.8A) continued to be called but less active true origin sites were excluded (ie more false negatives were generated).

This transition from exponentiality and non-specific site calling to linearity and true origin calling occurred at the p cut-off value of  $10^{-9}$ .

The same-analysis of replicates 2 (Fig.4.9B) and 3 (Fig.4.9C), showed the same pattern and possessed a switch from linearity to exponential function at approximately the same p cut-off value.

These data provided a rough objective indication that a p cut-off value of  $10^{-9}$  was appropriate for calling origins. To confirm this finding, I visualised the origins called in this p cut-off titration across the genome in IGV (Fig.4.10).



**Figure 4.10:** The IGV images show the mapped sequencing data generated from the replicated HL DNA (dark red for replicate 1 (A) or orange-red for replicate 3 (B)), the corresponding unreplicated LL DNA (grey for both replicates), the reference genes (purple) and the corresponding origins called by MACS peak caller at increasingly stringent p cut-off values, where  $10^{-5}$  was the most lenient/least stringent and  $10^{-11}$  was the most stringent, for replicates 1 (A; gradient of dark red colours) and 3 (B; gradient of orange-red colours) (for all of the replicates, the read numbers for HL and LL were 150 million reads (replicate 1 LL was ~140 million reads as this was the size of the original file)). The chromosome and position (bright red marker on the chromosome) on the chromosome are also indicated above the HL and LL profiles. The examples of called origins at different p cut-off values shown here are those found at the EZH2 promoter (i), LINC02005 gene body (ii) and the MYC promoter (iii) for both replicates 1 (A) and 3 (B). The selected p cut-off value of  $10^{-9}$  is highlighted in bold. The same IGV example images for replicate 2 are shown in appendix A4; Fig.A4.4A.

Examples of called origins at different p cut-off values in replicates 1 and 3 are shown in Figure 4.10. The highly active origin at the EZH2 promoter was called in both replicates at all p -cut-off values (Fig.4.10A). However, the width of the called sites reduced marginally at more stringent p cut-off values.

By contrast, a site in the gene body of LINC02005 (Fig.4.10B) was called at more lenient p cut-off values but not at more stringent ones, in both replicates. There was a slight enrichment in read accumulation in HL DNA but also a small enrichment in LL DNA. Thus, it seems less likely to be a true origin and it was correctly not called at the  $10^{-9}$  cut-off.

Finally, Fig.4.10C showed the broad replication initiation zone found at MYC. In both replicates, MACS called two neighbouring origins within the MYC broad zone for all p cut-off values. The key difference between the more and less stringent p cut-off values was in the width of the called sites; they became narrower as the p cut-off value became more stringent.

Replicate 2 (appendix A4; Fig.A4.4A/B/C) showed similar results as replicates 1 (Fig.4.10Ai/Bi/Ci) and 3 (Fig.4.10Aii/Bii/Cii).

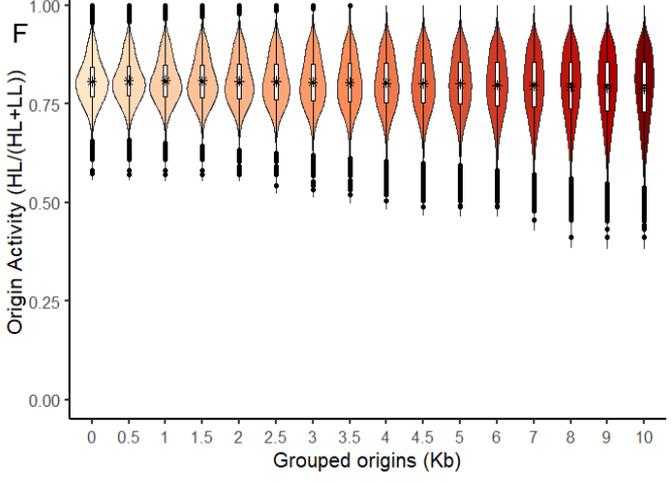
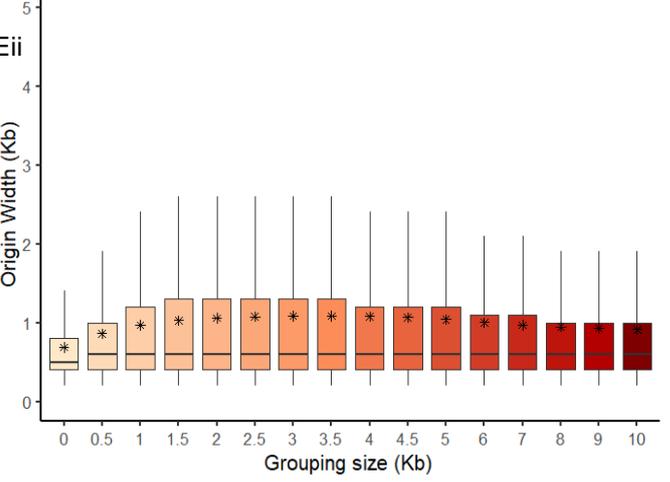
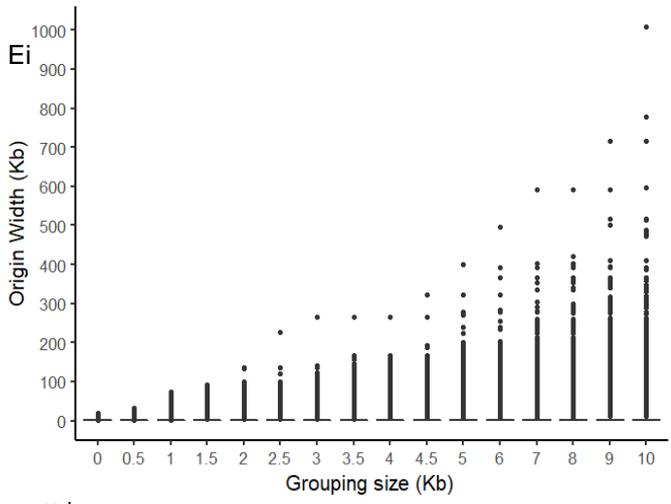
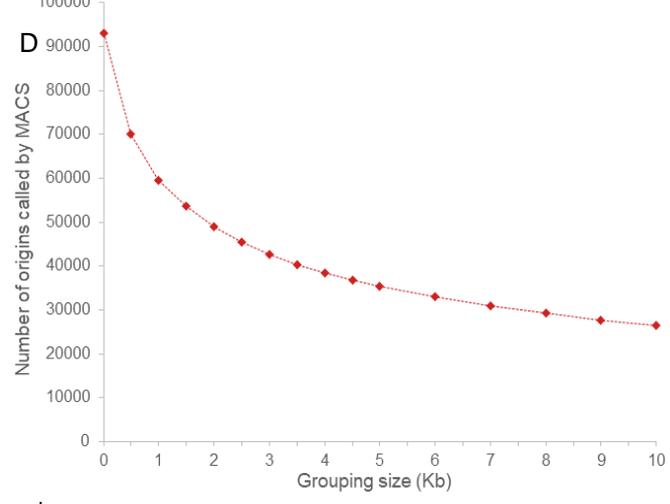
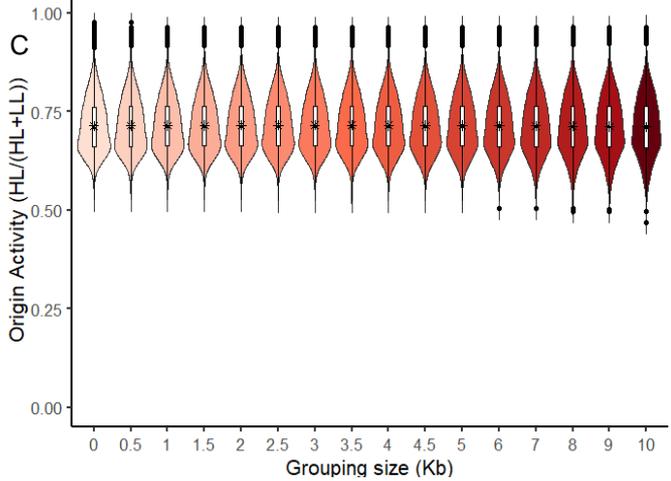
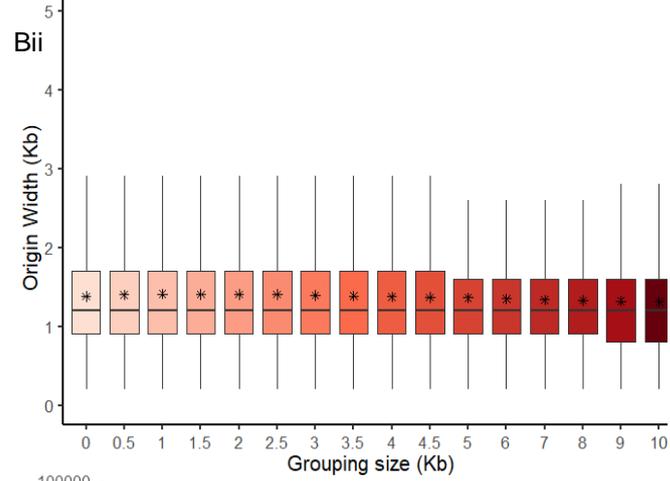
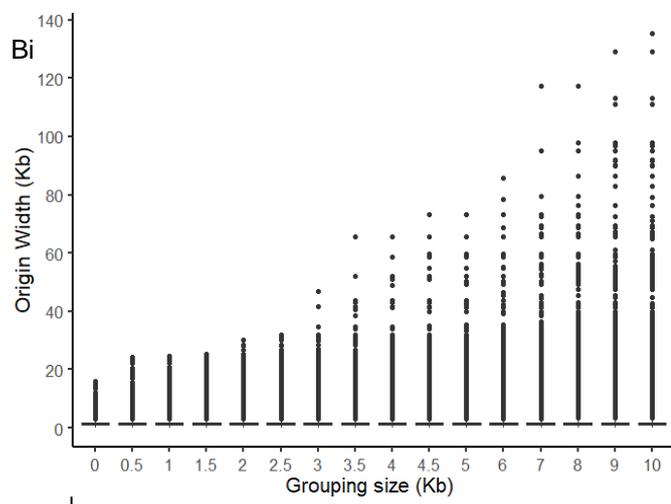
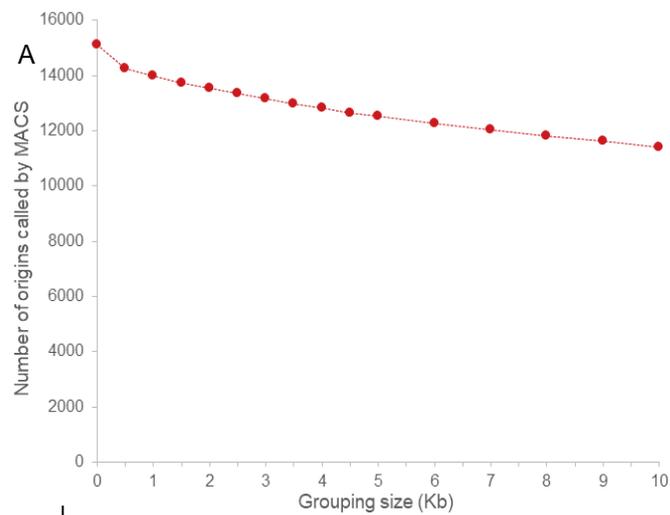
Fig.4.10 confirmed through visualisation of the sequenced data and MACS called sites that a p cut-off value of  $10^{-9}$  appeared to be appropriate for accurate replication origin/site calling.

From the data in Fig.4.9 and Fig.4.10, I used the p cut-off value of  $10^{-9}$  for MACS peak calling in all subsequent ds-iniSeq analyses.

### *3. Origin calling – at what distance should neighbouring origins be grouped?*

As the MYC replication initiation zone showed, there are wider regions of DNA replication initiation known as initiation zones (9,10), in addition to smaller discrete sites. MACS peak caller is better suited to smaller discrete sites than broader zones (8). Consequently, MACS may have called initiation zones as multiple smaller discrete sites. I therefore conducted a “grouping” titration, where I grouped together called replication initiation sites within a selection of distances of one another.

I selected grouping sizes from 500bp to 10Kb. For each grouping size, I determined the number of origins called by MACS at p cut-off value of  $10^{-9}$ , their widths and relative origin activity (Fig.4.11). Replicates 1 and 3 are shown here, as they possessed the smallest (15,105 ungrouped origins) and largest (129,410 ungrouped origins) number of called replication sites respectively. The same analysis conducted on replicate 2 is shown in appendix Fig.A4.5.



**Figure 4.11:** The numbers of origins called by MACS peak caller at a p cut-off value of  $10^{-9}$  (HL 150 million reads for all replicates, LL ~140 and 150 million reads for replicates 1 and 3 respectively), where origins were either ungrouped, or grouped where origins were within 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 6, 7, 8, 9 and 10Kb of one another (aka grouping size). (A) shows the number of origins at found with these grouping parameters for replicate 1. (B) shows the widths of the replicate 1 origins, at the grouping parameters, for both the overall widths for all origins (i) and the origins with widths of up to 5Kb (ii) to highlight the interquartile ranges, medians and means (highlighted with \*). (C) shows the origin activities of the replicate 1 origins, at the grouping parameters where the interquartile ranges, outliers and medians were indicated with an overlaying boxplot. The means are indicated with an \*. (D) shows the number of origins found with these grouping parameters for replicate 3. (E) shows the widths of the replicate 3 origins at the grouping parameters, for both the overall widths for all origins (i) and the origins with widths of up to 5Kb (ii) to highlight the interquartile ranges, medians and means (highlighted with \*). (F) shows the origin activities of the replicate 3 origins, at the grouping parameters where the interquartile ranges, outliers and medians are indicated with an overlaying boxplot. The means are indicated with an \*.

Replicate 1 (Fig.4.11A) showed a moderate drop in origin number from ungrouped origins to grouping size 500bp. From grouping sizes of 500bps to 10Kb, the origin number showed a gentle reduction with increasing grouping size.

The overall origin width (including outliers) (Fig.4.11Bi) increased as grouping size increased, but the median and mean widths (Fig.4.11Bii) remained fairly constant across all grouping sizes. This may indicate that most MACS called origins were unaffected by grouping, suggesting that they were smaller discrete sites. The relative origin activity generally declined as grouping size increased (Fig.4.11C). Additionally, the distribution of relative activities became more disperse with a larger range with increasing grouping size, indicating that more origins possessed lower relative origin activities.

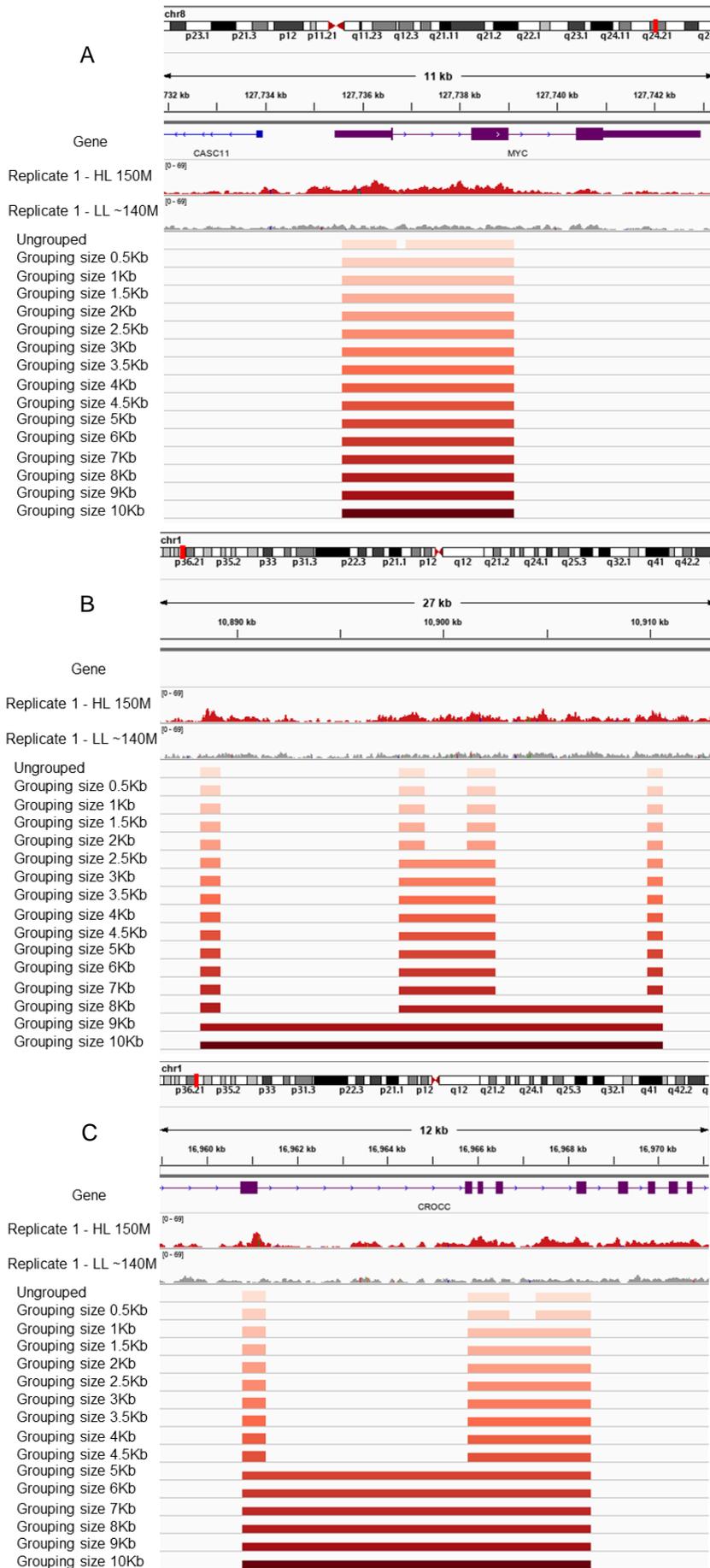
Replicate 3 (Fig.4.11D) showed a substantial reduction in origin number from ungrouped origins to grouping sizes of 4-5Kb. From grouping size 4-5Kb to 10Kb, the number of origins continued to reduce at a less substantial rate. The number of origins called in replicate 3 was much greater than those called in replicate 1 at all grouping sizes. The effect grouping size had on relative origin activities and their distributions was similar to, but more exaggerated than, that seen in replicate 1 (Fig.4.11F). The grouping size titration and impact on relative origin activity in replicate 2 (appendix Fig.A4.5A) showed an almost identical pattern to replicate 3.

The overall origin widths (including outliers) (Fig.4.11Ei) increased as grouping size increased. The mean width increased from the ungrouped origins to the grouping size of 1.5kb, remained roughly constant for grouping sizes 1.5-7Kb, and then declined marginally as grouping size increased to 10Kb (Fig.4.11Eii). The mean origin widths for grouping sizes 8-10Kb remained higher than that of ungrouped origins. The titration of grouping size in replicate 2 (appendix Fig.A4.5B) showed an almost identical pattern (replicate 2 had fewer

and shorter origins) to replicate 3. Again, these may suggest that most called origins were unaffected by grouping.

These data showed that grouping neighbouring origins affected origin number, overall width (outliers included) and relative origin activity. It appeared that this effect was more prominent in the replicates with greater origin numbers.

I then visualised the origin sites generated in the grouping titration for replicate 1 across the genome in IGV (Fig.4.12; replicate 2 & 3 examples shown in appendix Fig.A4.6&A4.7).

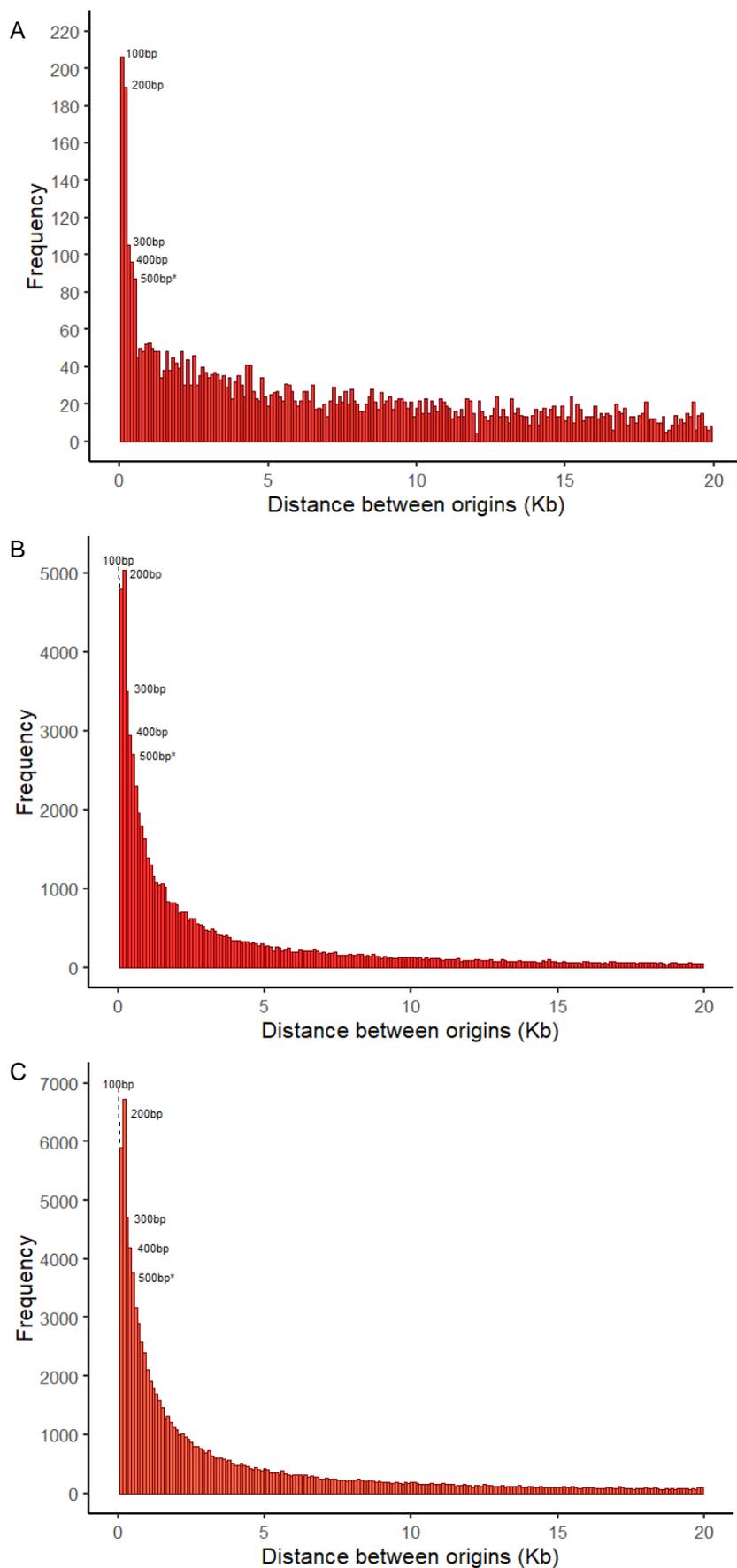


**Figure 4.12:** The IGV images show the mapped sequencing data generated from the replicate 1 replicated HL DNA (dark red), the replicate 1 unreplicated LL DNA (grey for all replicates) (HL 150 million reads, LL ~140 million reads) and the reference genes (purple). The chromosome and position (bright red marker on the chromosome) on the chromosome are also indicated above the HL and LL profiles. The origins (MACS peak caller,  $p = 10^{-9}$ ) identified from the grouping titration was also shown, from ungrouped origins to grouping sizes of 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 6, 7, 8, 9 and 10Kb (origins with those distances were grouped together and classed as 1 larger origin). (A) shows the example of the replication region/zone found at the MYC promoter, which demonstrates that grouping was required. (B) shows the example of origin(s) found at an intergenic region, which highlighted that large grouping sizes were inappropriate. (C) shows an example of origin(s) found at gene body region, which highlighted that grouping size greater than 0.5Kb was inappropriate.

The example of MYC (Fig.4.12A) demonstrated that grouping origins within 0.5Kbs of one another resolved the issue of two smaller neighbouring origins which should have been called as 1 wider replicating region/zone.

An example that demonstrated that grouping sizes greater than 2.5Kb were inappropriate is shown in Fig.4.12B. One further example (Fig.4.12C) showed that grouping sizes greater than 1Kb were inappropriate, as it grouped together two distinct replicating sites; the grouping size of 0.5Kb preserved this distinction.

To ultimately determine which grouping size to use, I examined the distances between all the ungrouped origins in each replicate (HL was fixed at 150M for all replicates, and LL was fixed at 150M for replicates 2 and 3, and ~140M for replicate 1). Fig.4.13 shows the histogram of these distances (100bp bins).



**Figure 4.13:** The frequencies of the distances between origins called by MACS peak caller (HL 150 million reads for all replicates, LL ~140 and 150 million reads for replicates 1 and 2/3 respectively), at a p cut-off value of  $10^{-9}$  (bins were 100bp). The distances of 0-100bp, 101-200bp, 201-300bp, 301-400bp, 401-500bp are indicated with “100bp”, “200bp”, “300bp”, “400bp” and “500bp” respectively. Replicates 1 (A), 2 (B) and (C) show a peak in frequencies at distances below 500bp. The grouping size of 500bp/0.5Kb was selected and is highlighted here with \*. The distances between origins shown here were between 0 and 20Kb for all replicates. The range of distances between origins were from 0kb to 30,083.6Kb, 217,821.7Kb and 26,662.7Kb for replicates 1 (A), 2 (B) and 3 (C) respectively.

Replicate 1 (Fig.4.13A) showed a large peak in potential origins within 1-100bps of one another. The frequency decreased rapidly to the “baseline” frequency at potential origins within 501-600bps of one another and remained at that “baseline” frequency to the end of the plot; frequencies of distances between origins remained at the “baseline” from 20Kb to the end of its range (data not shown). The higher frequency of neighbouring origins within smaller distances of one another indicated potential sites, such as MYC (Fig.4.11), were they should be classed as one wider origin/initiation zone. Therefore, I grouped origins within 500bps of one another.

Replicate 2 (Fig.4.13B) showed a large peak in smaller distances between origins, with the peak distance of 101-200bp. The frequency of distances between origins reduced drastically until ~4Kb and continued at this low frequency “baseline” until the end of the plot (20Kb) and beyond, to the end of its range (data not shown).

As with replicate 1, the distance between origins in replicate 2 showed a substantial reduction in frequency by 401-500bps, when compared to the height of the peak; a little under half that of the height of the peak. Therefore, grouping called sites within 500bps of one another was appropriate for replicate 2.

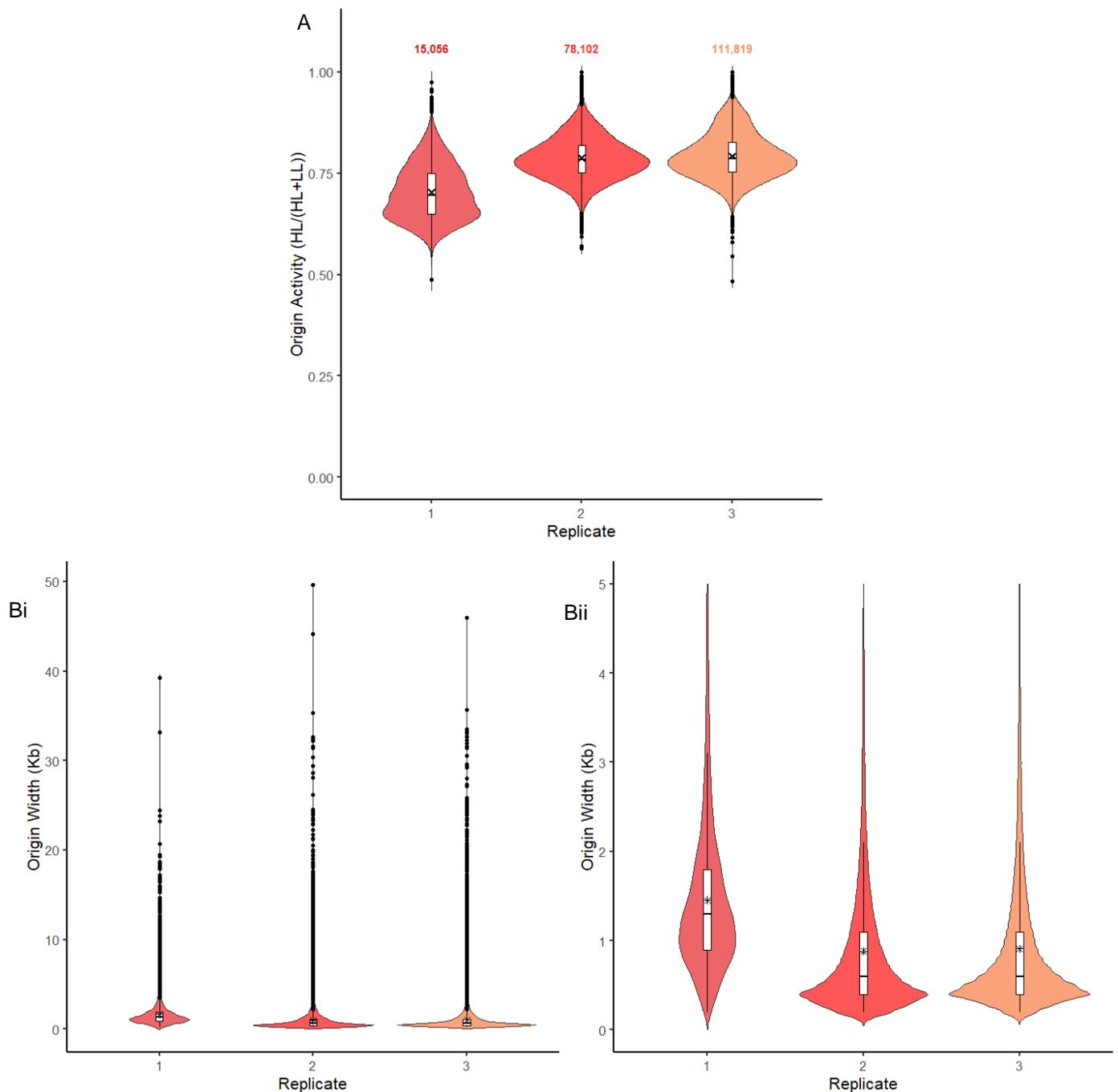
Replicate 3 (Fig.4.13C) showed an almost identical pattern to that of replicate 2, whereby the peak distance was 101-200bp and the frequency of distances between origins declined drastically until around 5Kb. As with replicate 2, the frequency of distances between origins in replicate 3 was roughly half that of the height of peak at 401-500bps. Therefore, grouping called sites within 500bps was also suitable for replicate 3.

Moreover, the absolute frequencies of the peaks were much greater in replicates 2 (~5000 at 101-200bps) and 3 (~6800 at 101-200bp), when compared to that of replicate 1 (~210 at 0-100bp). This indicated a larger number of origins closer to one another/in the same genomic area. In conjunction with the much larger number of origins, this observation may imply that the replicates 2 and 3 underwent a greater extent of replication and potentially represent a later time point in replication, than replicate 1.

From the analyses documented in Fig.4.7-13, I determined that the optimal conditions for MACS peak calling of replication origins from ds-iniSeq data included a p cut-off value of  $10^{-9}$  and grouping of origins within 500bps of one another. With these parameters in place, I performed MACS peak calling and read count quantification on all three replicates.

#### 4.2.6 ds-*ini*Seq replicates – called origins and overlap analysis

Using MACS, I generated a list of replication origins for replicates 1, 2 and 3. The origin number, their relative origin activities, and widths are shown in Fig.4.14.



**Figure 4.14:** (A) The overall files of replicates 1, 2 and 3 were subject to MACS peak calling, with a p cut-off value of  $10^{-9}$ , and called origins within 0.5Kb were grouped together. Origin activities were calculated for each origin. The total number of identified origins for each replicate are shown above the corresponding violin plots of origin activities. The interquartile ranges, medians, outliers and means (X) are also shown. (B) The widths of the origins identified in replicates 1, 2 and 3, for all origins (i) and the origins with widths of up to 5Kb (ii) to highlight the interquartile ranges, medians and means (highlighted with \*).

Replicate 1 had fewest origins (15,056), followed by replicate 2 (78,102). Replicate 3 had the largest number of origins (111,819) (Fig.4.14A). The relative activities of replicate 1 origins possessed the widest distribution. Replicates 2 and 3 possessed a highly similar distribution of relative origin activities, which were smaller and more active than the replicate 1 activities. The distribution of origin activities of replicate 3 showed a marginal secondary enrichment of origins with higher relative activities, which was not observed in replicates 1 or 2. One must remain careful when comparing these replicates as there were differences in the amount of DNA recovered and amount of potential contaminating LL DNA in each replicate. As previously stated, the origin activities are relative and as such the absolute values cannot be used to draw a conclusion but may be indicative of activities of these origins.

The higher relative origin activities of replicates 2 and 3 provided evidence suggesting that they may represent later “time points” in DNA replication, compared to replicate 1. This may have resulted from more efficient replication initiation reactions that progressed more quickly *in vitro*. Alternatively, replicates 2 and 3 may contain greater levels of contaminating S-phase nuclei than replicate 1. I minimised the extent of S-phase contamination by ensuring that there was <5% S-phase nuclei in the nuclei preparations used in these ds-iniSeq experiments (assessed via immunofluorescence). One further control to address the extent of S-phase contamination within these experiments is to carry out a buffer only control, whereby the replication reaction of the ds-iniSeq experiment is conducted in the absence of cytosolic extract; only S-phase nuclei would be able to replicate. In future, one should consider assessing the degree of S-phase contaminating nuclei using flow cytometry prior to use in ds-iniSeq experiments, in addition to confocal microscopy.

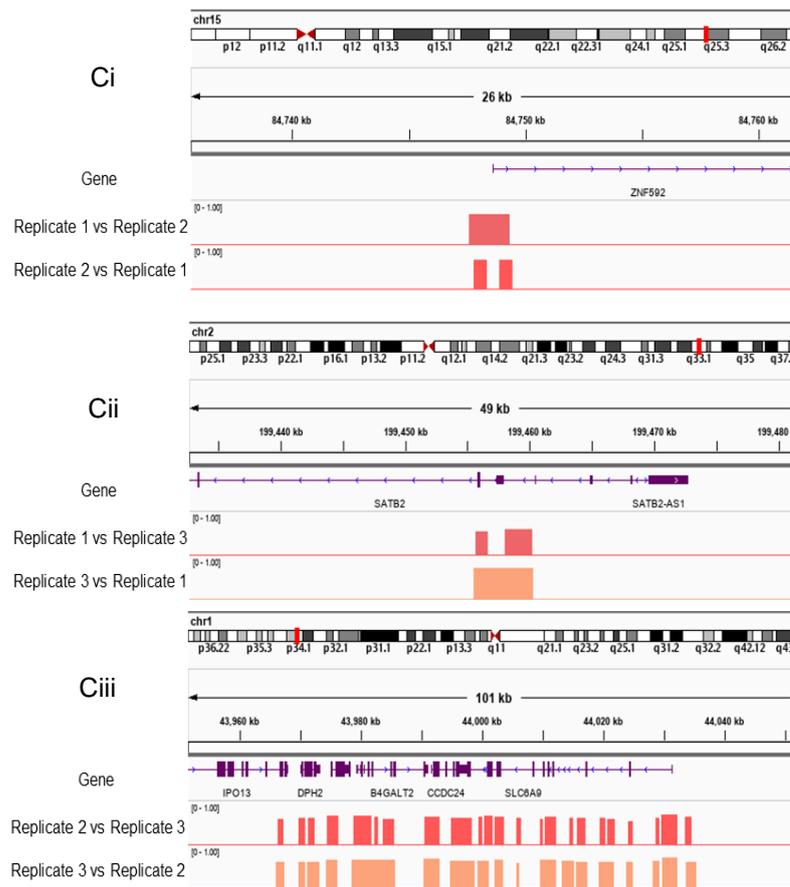
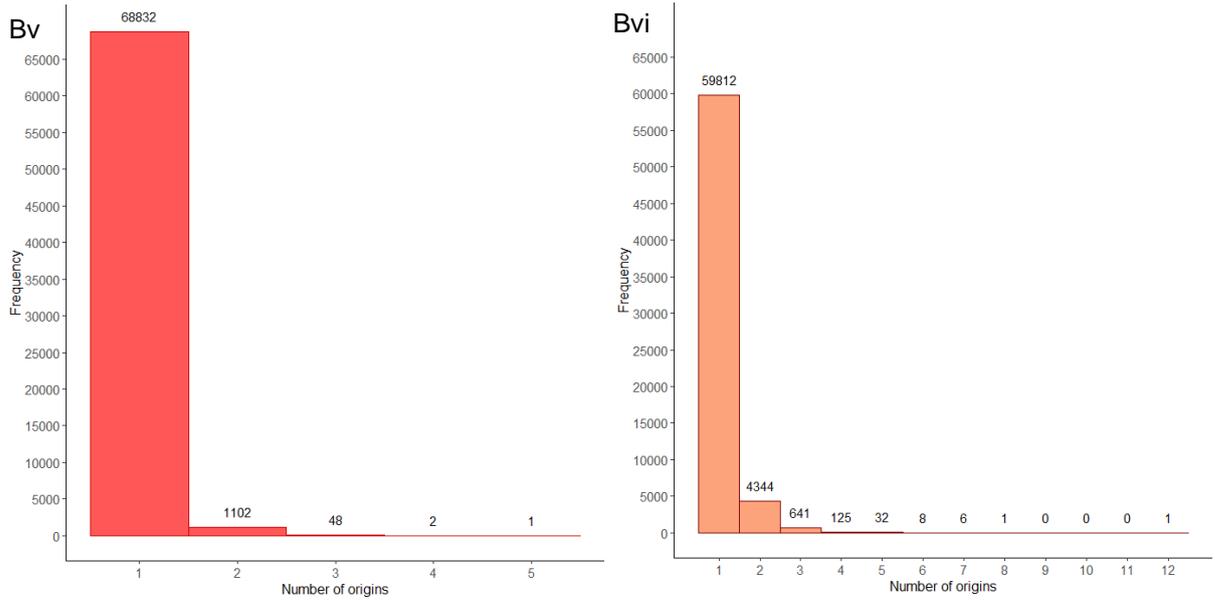
The range of widths of origins from replicate 2 was greater than that of replicate 3, which was greater than that of replicate 1. However, the mean width of replicate 1 origins was greater than that of replicates 2 and 3. The widths of origins of replicate 1 tended to a normal distribution at sizes between 0 and 3kb, with a mild skew to smaller widths, whereas the widths in the same region for replicates 2 and 3 were highly skewed to the smaller widths, with peaks around 200-400bp.

Overall, I was able to successfully identify origins in all three replicates using the parameters elucidated in figures 4.7-4.13 (section 4.2).

I then performed an overlap analysis of the origins called in all three replicates, to identify those origins which were common between replicates (Fig.4.15).



**Figure 4.15:** The number of overlapping origins between pairs of replicates: (Ai) origin overlap of replicates 1 & 2; (Aii) origin overlaps of replicates 1 & 3; (Aiii) origin overlaps of replicates 2 & 3. The top value shows the origin number of the replicate on the left ((Ai) = 2; (Aii) = 3; (Aiii) = 3) with those of the replicates on the right ((Ai) = 1; (Aii) = 1; (Aiii) = 2); the bottom value shows the inverse. (B) Frequencies of origins overlapping with 1 or more origin in the opposing replicate. Overlapping origin conditions were: (i) replicate 1 with replicate 2, (ii) replicate 2 origins with replicate 1 (iii) replicate 1 origins with replicate 3, (iv) replicate 3 origins with replicate 1.



(B) Frequencies of origins overlapping with 1 or more origin in the opposing replicate. Overlapping origin conditions were: (v) replicate 2 origins with replicate 3, and (vi) replicate 3 origins with replicate 2. (C) IGV images of origins that overlapped with more than one origin, where (i) shows 2 replicate 2 origins overlapping with 1 replicate 1 origin, (ii) shows 2 replicate 1 origins overlapping with 1 replicate 3 origin, and (iii) shows 3 and 2 replicate 2 origins overlapping with separate single replicate 3 origins.

The Venn diagrams showed the number of origins that overlapped in replicate 1 and 2 (Fig.4.15Ai), replicates 1 and 3 (Fig.4.15Aii) and replicates 2 and 3 (Fig.4.15Aiii).

There was an extremely high overlap of origins found in replicate 1 (replicate 1 origins), with those in replicates 2 (4.17Ai) and 3 (4.17Aii); >93% replicate 1 origins overlapped with those in replicates 2 and 3. The origins found in replicate 2 (replicate 2 origins) and replicate 3 (replicate 3 origins) also showed a high concordance with one another (Fig.4.15Aiii), but not to the same extent as overlaps with replicate 1; ~60% of replicate 3 origins overlapped with ~90% of replicate 2 origins.

Two values were generated for each overlap analysis, which was as a consequence of the Genomic Ranges software that was used to generate these values. For Fig.4.15Ai, the top value represented the number replicate 2 origins overlapping replicate 1 origins. The bottom value represented the number of replicate 1 origins overlapping replicate 2 origins. More replicate 2 origins (14,399) overlapped with replicate 1 origins than the reverse analysis (14,142).

For Fig.4.15Aii, the top value represented the number of replicate 3 origins overlapping replicate 1 origins, and the bottom value was the reverse. More replicate 1 origins (14,825) overlapped with replicate 3 origins, than the reverse analysis (14,372).

Finally, the top value in Fig.4.15Aiii showed the number of replicate 3 origins overlapping replicate 2 origins, and the bottom value was the reverse. More replicate 2 origins (69,985) overlapped with replicate 3 origins, than the reverse analysis (64,970).

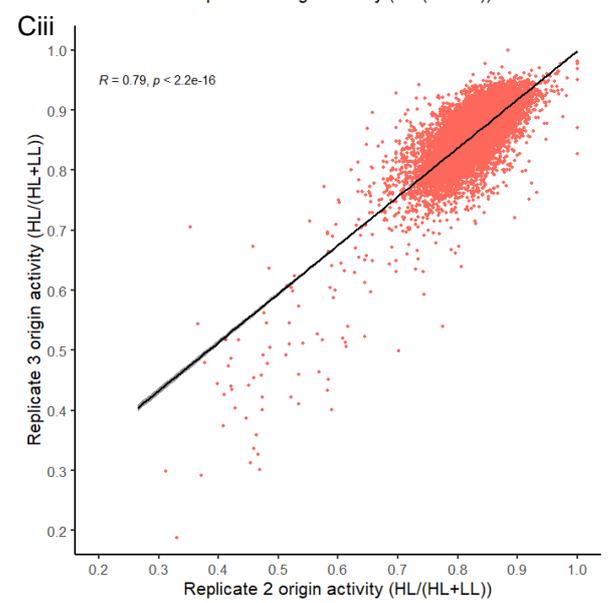
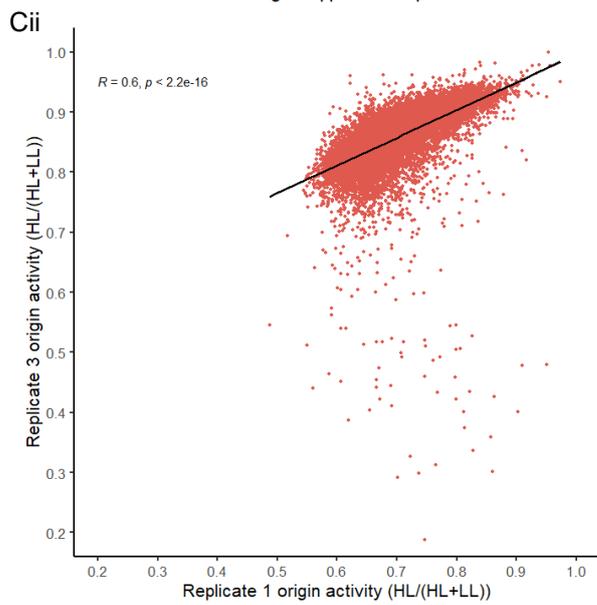
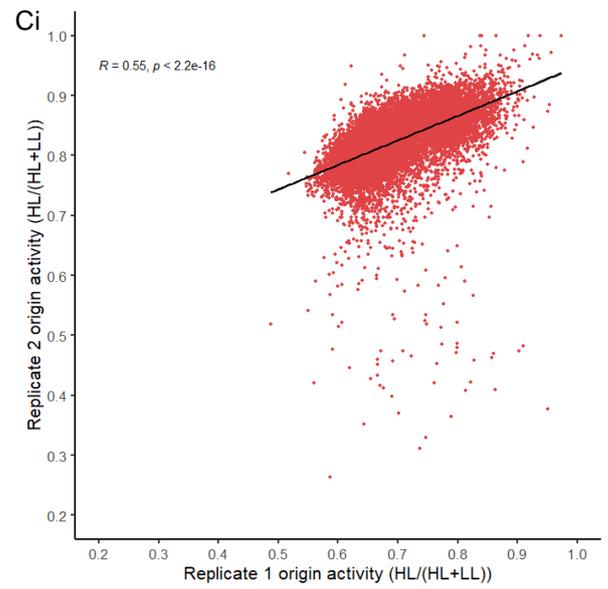
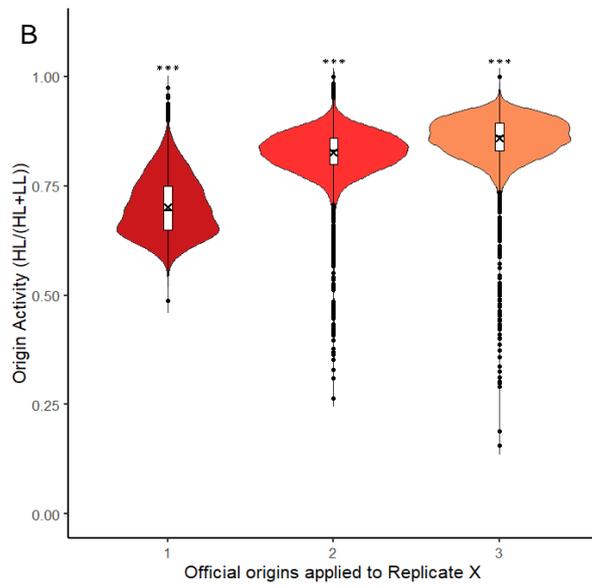
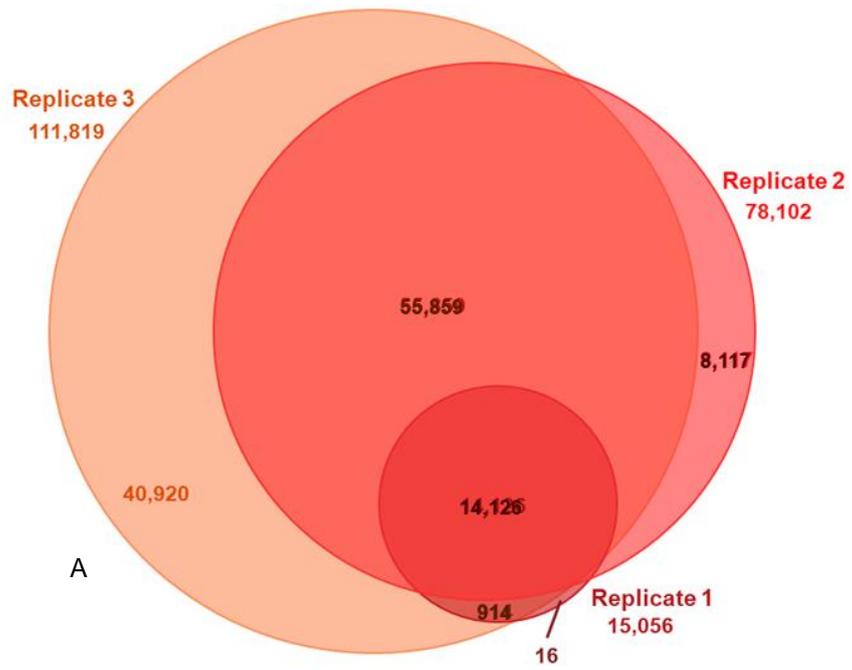
There was a clear discrepancy in the number of origins that overlapped between each set of replicates, which was dependent on the replicate being analysed (Fig.4.15A). It is possible that the higher overlap value for each overlap analysis could indicate a scenario where multiple origins in one replicate overlapped with one larger origin in the opposing replicate.

To test this hypothesis, I assessed how frequently origins of one replicate overlapped with multiple origins of the opposing replicate (Fig.4.15B). More replicate 1 origins overlapped with 2 or more replicate 2 origins (Fig.4.15Bi; 579 origins), than replicate 2 origins overlapped with 2 or more replicate 1 origins (Fig.4.15Bii; 327 origins). Whereas the number of replicate 1 origins that overlapped with 2 or more replicate 3 origins (Fig.4.15Biii; 81 origins) was much lower than the number of replicate 3 origins that overlapped with 2 or more replicate 1 origins (Fig.4.15Biv; 445 origins). Similarly, the number of replicate 2 origins that overlapped with 2 or more replicate 3 origins (Fig.4.15Bv; 1,153 origins) was much lower than the number of replicate 3 origins that overlapped with 2 or more replicate 2 origins (Fig.4.15Bvi; 5,158 origins, including 1 origin overlapping 12 replicate 2 origins).

Examples of these origin overlap scenarios were visualised on IGV. These IGV images (Fig.4.15C) showed (i) two replicate 2 origins overlapping one larger replicate 1 origin, (ii) two smaller replicate 1 origins overlapping one larger replicate 3 origins, and (iii) a selection of origins that overlapped between replicates 2 and 3, including instances where three smaller replicate 2 origins overlapped with one much larger replicate 3 origin, and two smaller replicate 2 origins overlapped with one replicate 3 origin.

Taken together, these data support a scenario where multiple origins in one replicate overlapped with one larger origin in the opposing replicate.

Finally, I performed a 3-way overlap analysis between all 3 replicates, in order to assess the concordance between all three replicates and identify a list of origins that were consistently present in all replicates, that I was able to use for any subsequent analyses (Fig.4.16).



**Figure 4.16:** (A) The overlap analyses of origins identified in replicates 1, 2 and 3. These origins were identified through MACS peak calling ( $p=10^{-9}$ ) and origins within 0.5Kb were grouped together. The overlap values here show the number of replicate 1 origins that overlapped with replicates 2 and 3. The overlap value for replicates 2 and 3, show the number of replicate 2 origins that overlapped with replicate 3. The 14,126 replicate 1 origins that overlapped with replicates 2 and 3, known as the “official origins” list, were applied to replicates 1, 2 and 3 independently to determine the origin activity of the common origin sites. (B) displays the origin activities of these sites for all three replicates, and the interquartile ranges, medians, outliers and means (X) are indicated. These data sets did not possess equal variance and a Welch’s test was performed to determine significance. The following Welch’s tests were performed; replicate 1 vs replicate 2, replicate 1 vs replicate 3, and replicate 2 vs replicate 3. All tests returned a p value of  $p < 2 \times 10^{-16}$  which demonstrated significant differences between all the origin activities of all replicate combinations (\*\*\*) indicated  $p < 0.001$ ). (C) The scatter of the origin activities (official origins applied to each replicate) of the following pairings: (i) replicate 1 (x-axis) vs replicate 2 (y axis); replicate 1 (x-axis) vs replicate 3 (y-axis); replicate 2 (x-axis) vs replicate 3 (y-axis). A linear regression, Pearson test for correlation and ANOVA test for significance were conducted for each pairing. The correlational R value and the ANOVA test result are indicated on the plot.

The Venn diagram of the overlap analyses between all 3 replicates (Fig.4.16A) showed that the origins called in replicate 1 almost completely overlapped (94%) with the origins called in both replicates 2 and 3. From this analysis, I was able to collate a list of the 14,126 replicate 1 origins that overlapped with replicates 2 and 3, now named the “official origins” list.

I then applied this “official origins” list to each replicate individually, to investigate the relative origin activities of these replicate 1 origin sites common to all three replicates (Fig.4.16B). These “official origins” sites provided the genomic positions of the sites where read counts were quantified using the SeqMonk feature quantitation tool; as such they acted as a mask for quantification of read count measurements. The relative activities of these sites were calculated as before, via a ratio of the localised read count accumulation of HL DNA to the total read count of both HL and LL DNA.

The “official origins” of replicate 1 possessed the lower relative activities with the widest distribution of relative activities (Fig.4.16B). The relative origin activities the “official origins” of replicate 2 were much greater and had a much narrower distribution, than that of replicate 1. The distribution of the relative origin activities of the “official origins” in replicate 3 was similar to that of replicate 2. The distributions of the relative activities for each replicate of these “official origins” reflect the data shown in Fig.4.14.

I performed statistical analyses of these origin activity values. As these data were normally distributed, but did not possess equal variance, I used the non-parametric Welch’s test between pairs of replicates (ie replicate 1 vs 2, replicate 1 vs 3, replicate 2 vs 3). I found that these were all highly significantly different from one another ( $p < 2 \times 10^{-16}$  for each pair of analyses). Therefore, I treated the data for each replicate as independent of one another and predominantly used the replicate 1 origin data for subsequent analyses.

I also performed a linear regression (as there were differing amounts of DNA recovered and possibly differing degrees of LL contamination for each replicate) of the relative origin activities of the official origins in replicate 1 vs replicate 2 (Fig.4.16Ci); replicate 1 vs replicate 3 (Fig.4.16Cii); and replicate 2 vs replicate 3 (Fig.4.16Ciii). In all three comparisons there was a moderate to strong positive correlation which was highly significant. These indicate that those "official origins" that were highly active/have a high probability of firing in replicate 1 were also likely to be highly active/have a high probability of firing in replicates 2 and 3; the same was true for replicates 2 and 3.

#### 4.2.7 Conclusions

This chapter has demonstrated that the novel ds-iniSeq method was able to successfully and reliably produce and separate replicated DNA and unreplicated DNA from standard *in vitro* replication reactions. I was able to utilise the well-established MACS peak caller to identify origin sites. Furthermore, I demonstrated that the relative origin activity of called origins could be determined, using the "read count quantitation" tool in the SeqMonk software.

With these in place, I was able to successfully call and determine the relative activities of replication origins for three experimental replicates.

There was a discrepancy between the numbers of called origins and their origin activities for each of the three replicates. The absolute numbers of called origins and their relative origin activities of replicates 2 and, to a greater extent, 3 were much higher than that of replicate 1. Nearly all the replicate 1 origins overlapped with origins found in both replicates 2 and 3. This degree of overlap was much higher than the concordance of the two iniSeq replicates, which was around 54% (1). However, the inverse (replicates 2 and 3 origins overlapping replicate 1 origins) overlap was much lower than observed with the two iniSeq replicates(1), as there were far greater numbers of replicate 2 and 3 origins.

It is possible that replicates 2 and 3 represented a later "time point" in DNA replication than replicate 1, which would explain why almost all replicate 1 origins were present in all three replicates. I hypothesise that the replicate 1 origins represented the earliest stage of DNA replication of those observed here and would predominantly use the replicate 1 origins (that overlapped replicates 2 and 3 origins) for any subsequent analyses.

The difference in "time point" of these replicates may be explained by the biological variability between replication reactions of each replicate, whereby the initial *in vitro* replication reactions for replicates 2 and 3 progressed more efficiently/faster than that of replicate 1. This variability was most probably due to biological variability of the nuclear preparations; differences in the efficiency of prepared nuclei have been commonly seen in the human cell-

free system (indicated by incorporation of a fluorescently labelled tag across a 3-hour replication reaction).

Moreover, contamination of the late G1-phase template nuclei preparations with S-phase nuclei (that escaped synchronisation) would have contributed to higher background signal in the HL samples of the sequencing data. The impact of S-phase contamination was minimised as much as possible and nuclear preparations with <5% S-phase contamination (determined using the human cell-free system in the absence of cytosolic extract) were used for any ds-iniSeq experiment. A further experiment that might address this issue would be to conduct a ds-iniSeq experiment in the absence of cytosol (only S-phase nuclei would replicate and incorporate the BrdUTP). Should sufficient HL DNA be obtainable, this could offer a control for the effect of contamination by S-phase nuclei. Alternatively, one could assess S-phase contaminants through flow cytometry.

Overall, ds-iniSeq has resulted in a useable list of origin sites (replicate 1 origins overlapping replicates 2 and 3 origins) that I can employ to further investigate the specification and activation of human replication origins. The generation of relative origin activities will allow me to investigate the correlation between origin activity and genomic and epigenetic features previously associated with DNA replication. The *in vitro* replication reaction at the core of the ds-iniSeq method can be manipulated providing me with the unique opportunity to assess the impact of various replication factors on replication origin specification and activation.

## **Chapter 5: The analysis of the standard density-substitution initiation-site sequencing (ds-iniSeq) experiment**

### **5.1 Introduction**

In the previous chapter, I described the successful development of the standard ds-iniSeq protocol and the overlap analyses of three replicates generated from these reactions.

From these data, I have generated an “official” list of 14,126 origins consisting of replicate 1 origins that overlap with origins present in replicates 2 and 3. For the subsequent analyses in this chapter, these “official” origins will be referred to as ds-iniSeq origins. For this chapter, I generated a corresponding list of random sites (ds-iniSeq RS), which was proportional to the number of origins per chromosome; the relative activities for these random sites were obtained from the replicate 1 data. These were subsequently used for comparative analysis with the ds-iniSeq origins.

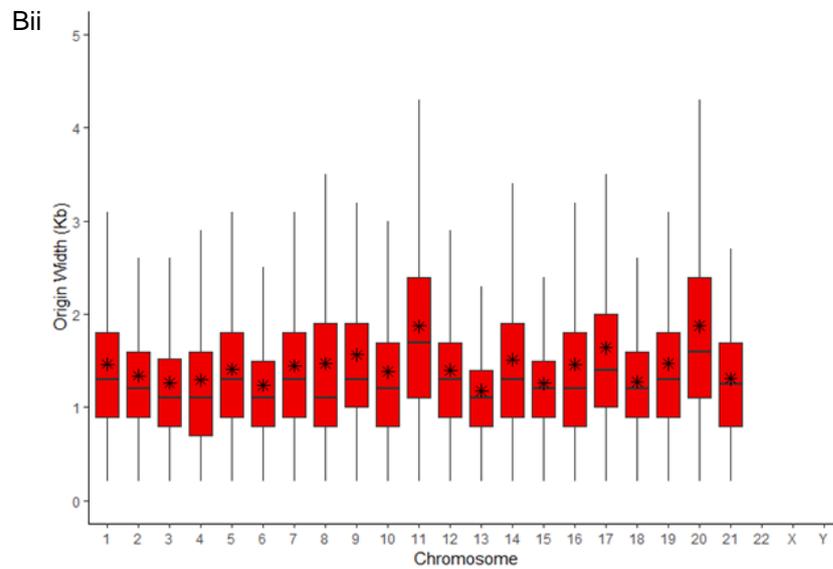
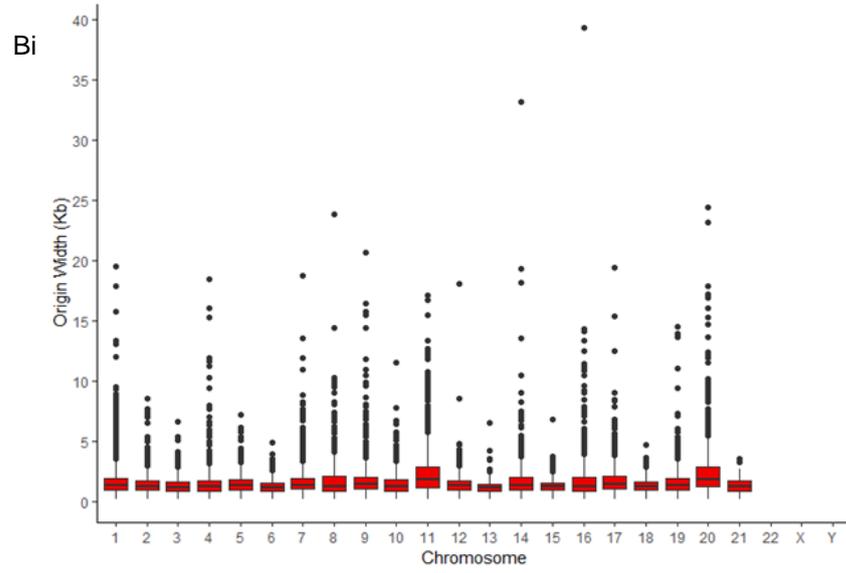
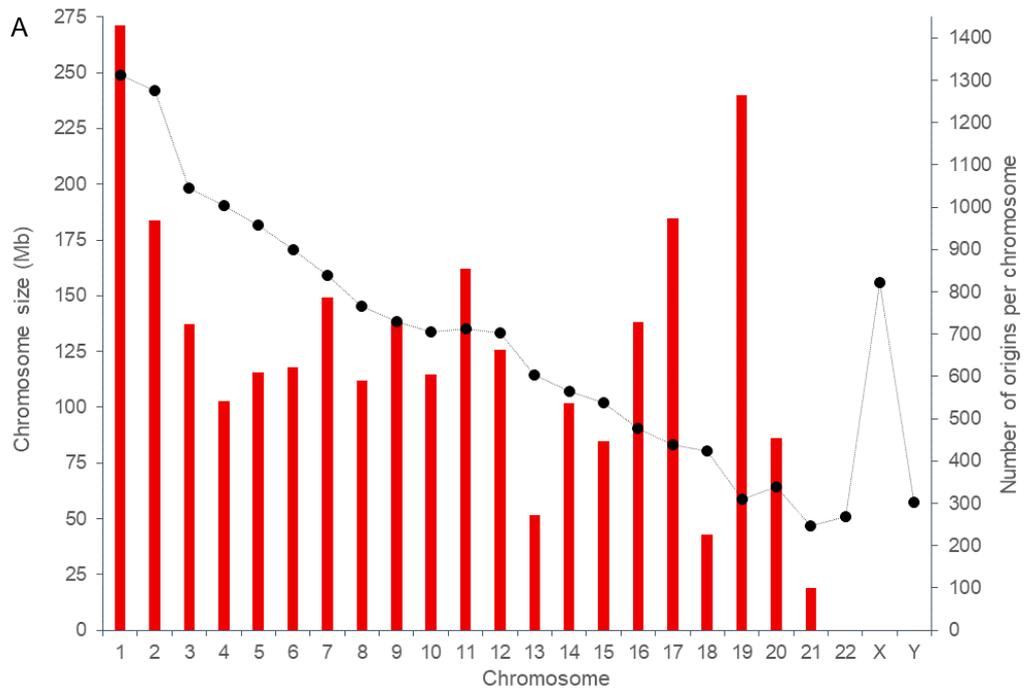
In chapter 4, I proposed that the reactions in replicates 2 and 3 represented a later timepoint in DNA replication. For this chapter, I have generated a list of replicate 2 origins that overlap with origins present in replicate 3 but not replicate 1, which will be referred to as ds-iniSeq23 origins. I also generated a list of random sites (ds-iniSeq RS23), which was proportional to the number of origins per chromosome; the relative activities for these random sites were obtained from the replicate 2 data. The ds-iniSeqRS23 and ds-iniSeq23 origins were used for analysis with replication timing data.

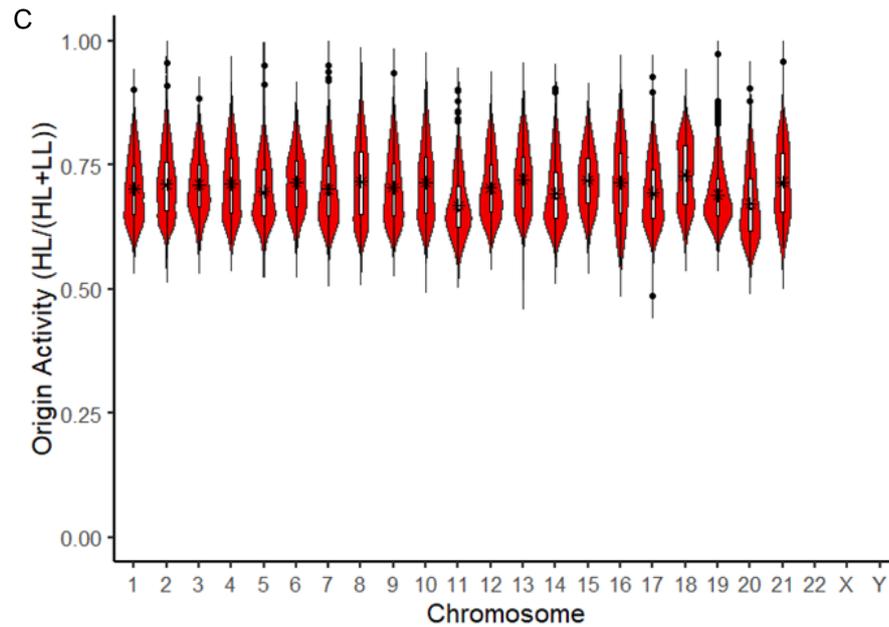
In this chapter, I describe and discuss the analysis of these sequencing data, including the make-up and distribution of the ds-iniSeq origins and how these compare to genomic and epigenetic features, and replication timing profiles.

### **5.2 Results and discussion**

#### *5.2.1 Origins per chromosome*

To assess the distribution of origins, their widths and relative activities between chromosomes, the ds-iniSeq origins were separated by chromosome and compared to each chromosome size (Fig.5.1).





**Figure 5.1:** The called origins found in replicate 1, that overlap with the called origins in replicates 2 and 3, were divided to produce the number of origins per chromosome. (A) The number of origins per chromosome (red bar chart on secondary Y axis) were compared to the size (Mb) of each chromosome (black scatter on primary Y axis). (B) The range of origin widths (Kb) per chromosome showing (i) all origins (up to 40Kb – outliers shown) and (ii) origins up to 5Kb in width (no outliers). The median width of all origins was 1.299Kb. (C) The origin activity of the origins per chromosome. For (B) and (C), the means were indicated with an \*.

Comparison of the number of origins per chromosome (Fig.5.1A) to the chromosome size (Mb) demonstrated that the number of origins on each chromosome did not follow the same trend as the chromosome size, making it unlikely that chromosome size influenced origin number, which was consistent with the findings of iniSeq (1) and SNS-seq (2).

The overall widths and interquartile ranges (IRs) (including outliers) of the origins per chromosome (Fig.5.1B) showed that they varied for each chromosome, but this did not appear to correlate with the chromosome size (Fig.5.1A; primary Y axis). Most of the origin widths on each chromosome did not differ excessively except for chromosome 11 and 20 origins, which extended to larger widths. The variation of ds-iniSeq origin widths was greater than origins identified by iniSeq (iniSeq origins) (1). The median width of the whole ds-iniSeq data set was 1299bp, which was only slightly larger than the median width of the iniSeq origins (1184bp) (1).

Origin width indicates the level of DNA replication during the replication reaction. The wider origins may have resulted from; a greater number of earlier firing origins, a higher replication fork progression rate, fusing of two neighbouring replication forks or a mixture of the three. Conversely the smaller ds-iniSeq origins could have resulted from fewer early firing origins and/or slower replication fork progression.

The relative origin activity of the origins on each individual chromosome (Fig.5.1C) also varied. There was no clear correlation between relative origin activity and chromosome size or numbers of origins per chromosome. There was a weak negative correlation between width and activity of origins genome-wide ( $r=-0.2$ ; appendix Fig.A5.1) which may indicate the extent of elongation from each origin (ie fork progression) or the potential effect of two converging neighbouring origins that have fused following elongation.

### *5.2.2 Comparison to alternative origin identification methods*

To assess the concordance between the ds-iniSeq origins with origins identified by other NGS methods, I performed an overlap analysis with a comparative analysis of the ds-iniSeq random sites (RS). I performed an overlap analysis of the ds-iniSeq origins (and RS) with the previous iniSeq origins (Fig.5.2) (1) and with origins identified by SNS-seq (SNS-seq origins) (Fig.5.3) performed by L. Koch-Lerner and G. Guilbaud (GG) (J. Sale group, LMB-MRC, Cambridge); both were performed on the EJ30 cell line.

#### *5.2.2a iniSeq overlap*



The ds-*iniSeq* origins and the *iniSeq* origins were visualised on the human genome, at the region of DNA near the *EZH2* gene, in IGV (Fig.5.2A). The image showed fewer ds-*iniSeq* origins than *iniSeq* origins at this DNA region and all the ds-*iniSeq* origins overlapped with the *iniSeq* origins. This was reflected genome-wide (Fig.5.2Bi) where ~60% of the 14,126 ds-*iniSeq* origins overlapped with 36% of the 25,385 *iniSeq* origins. By contrast, only 428 ds-*iniSeq* RS (~3%) overlapped with 455 (~3%) *iniSeq* origins (Fig.5.2Ci). This demonstrated an enrichment of ds-*iniSeq* origins at *iniSeq* origins sites indicating that the majority of ds-*iniSeq* origins that overlapped the *iniSeq* origins did not do so by chance; both were specifically located at the same sites.

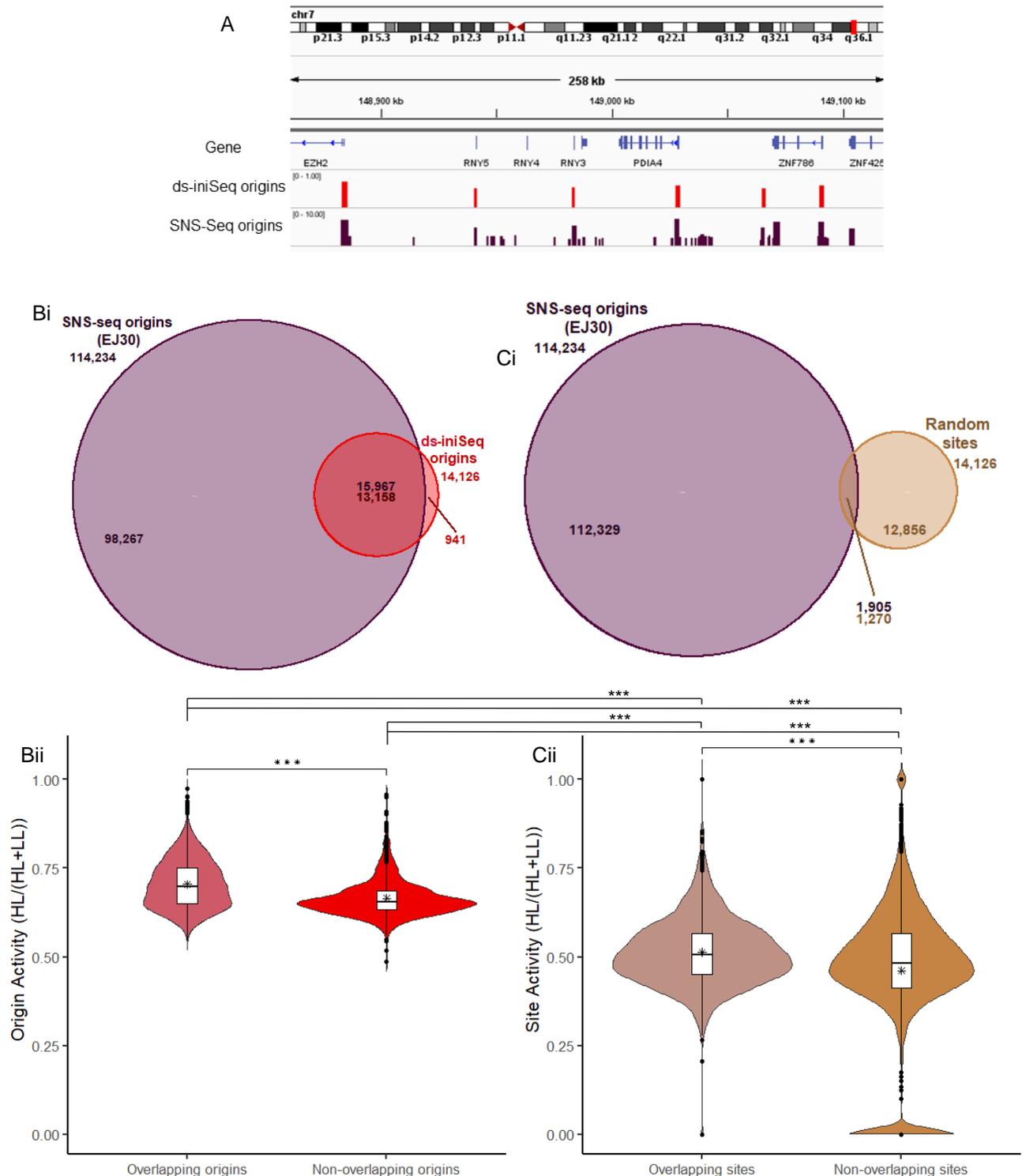
These ds-*iniSeq* origins showed an overall concordance of ~57% with *iniSeq* origins, which was better than the concordance of *iniSeq* with all other methods it was compared to (1). One potential explanation for this may be differences in origin specification and activation that arise from different cell lineages; ds-*iniSeq* and *iniSeq* were conducted on EJ30 cells, whereas *iniSeq* was compared with SNS-, OK- and bubble-seq data conducted on HeLa and GM06990 cells (1).

As ds-*iniSeq* offered the unique opportunity to assess relative origin activity levels, I assessed the difference in ds-*iniSeq* origins' activities that did and did not overlap with the *iniSeq* origins (Fig.5.2Bii). The relative activities of the ds-*iniSeq* origins that overlapped with *iniSeq* origins were significantly higher than those ds-*iniSeq* origins that did not overlap. The distribution of non-overlapping ds-*iniSeq* origins showed lower relative origin activities compared to those ds-*iniSeq* origins that do overlap.

This observation suggested that the origins that fire/initiate more consistently (ie appeared in more than one method of origin identification) were more active as they possessed a significantly higher relative origin activity, implying that the higher the relative activity, the more likely the ds-*iniSeq* origin was to initiate. As such, the origin activity acted as a proxy for the probability that a ds-*iniSeq* origin was to fire/initiate. Consequently, this support the stochastic model for origin firing (3,4) where origin firing is based upon probability, rather than a strictly defined firing timing pattern.

The relative activities of ds-*iniSeq* RS that overlapped with *iniSeq* origins were significantly higher than those ds-*iniSeq* RS that did not overlap (Fig.5.2Cii). *IniSeq* origins were detected because they possessed DNA replication activity; therefore, the randomly allocated sites overlapping them would, consequently, have higher activities. The relative activities of the ds-*iniSeq* origins that did and did not overlap with *iniSeq* origins were significantly higher than the ds-*iniSeq* RS that both overlapped and did not overlap with *iniSeq* origins.

### *5.2.2b SNS-seq overlap*



**Figure 5.3:** The called origins found in replicate 1 that overlap with the called origins in replicates 2 and 3, were compared to the called origins of SNS-seq conducted on the EJ30 cell line (L. Kock-Lerner and G. Guillaud (J.Sale group, MRC-LMC, Cambridge)). (A) An IGV image of the ds-iniSeq (red) and SNS-seq (plum) origins at the region near the EZH2 gene. (Bi) The number of ds-iniSeq origins that overlap with SNS-seq origins (red) and vice versa (plum). (Bii) The origin activity of the ds-iniSeq origins that do (overlapping origins) and do not (non-overlapping origins) overlap with SNS-seq origins. The mean is indicated with an \* and a student's T-test was conducted; \*\*\* indicates  $p < 0.001$ . (Ci) The number of ds-iniSeq random sites that overlap with SNS-seq origins (ochre) and vice versa (plum). (Cii) The origin activity of the ds-iniSeq random sites that do (overlapping sites) and do not (non-overlapping sites) overlap with SNS-seq origins. The mean is indicated with an \* and a student's T-test was conducted; \*\*\* indicates  $p < 0.001$ . An ANOVA and subsequent Tukey's post-hoc test was performed to assess significance between ds-iniSeq origins (Bii) and their corresponding random sites (Cii); the Tukey's test results are shown on the plot; \*\*\* indicates  $p < 0.001$ .

To compare origins identified by ds-iniSeq and SNS-seq (EJ30) in the same cell line, the corresponding origins were visualised in IGV (Fig.5.3A). There were fewer ds-iniSeq origins than SNS-seq origins at this DNA region and all of the ds-iniSeq origins overlapped with SNS-seq origins.

The overlap analysis of the ds-iniSeq and SNS-seq origins (Fig.5.3Bi) revealed a very high concordance; >93% of ds-iniSeq origins overlapped with ~14% of the 114,234 SNS-seq origins, which was far greater than that of any other comparative NGS origin identifying methods (1). The greater number of SNS-seq origins than ds-iniSeq origins may have been due to SNS-seq's better efficiency for detecting later initiating origins and/or a higher false positive rate resulting from incomplete lambda exonuclease digestion (2,5,6).

Comparatively, only 9% of ds-iniSeq RS overlapped with ~1.7% of the SNS-seq origins (Fig.5.3Ci), demonstrating an enrichment of >84% of ds-iniSeq origins at SNS-seq origins sites. This indicated that ds-iniSeq origins were specifically located at the same sites as SNS-seq origins. The concordance of ds-iniSeq origins with SNS-seq origins was much greater than that of iniSeq with SNS-seq origins (56%) (1). Comparison of these concordances was limited as the ds-iniSeq origins were compared to SNS-seq conducted on EJ30 cells, whereas the iniSeq origins were compared to SNS-seq conducted on HeLa cell, and there may have been cell line differences. However, comparison of iniSeq origins with the EJ30 SNS-seq data (data not shown), showed that 52% of the iniSeq origins overlapped with 17% of EJ30 SNS-seq origins. Therefore, the difference between concordance of ds-iniSeq (93%) and iniSeq (52%) origins with SNS-seq origins was retained even when the same cell lines were used for both methods and was not due to cell lineage differences.

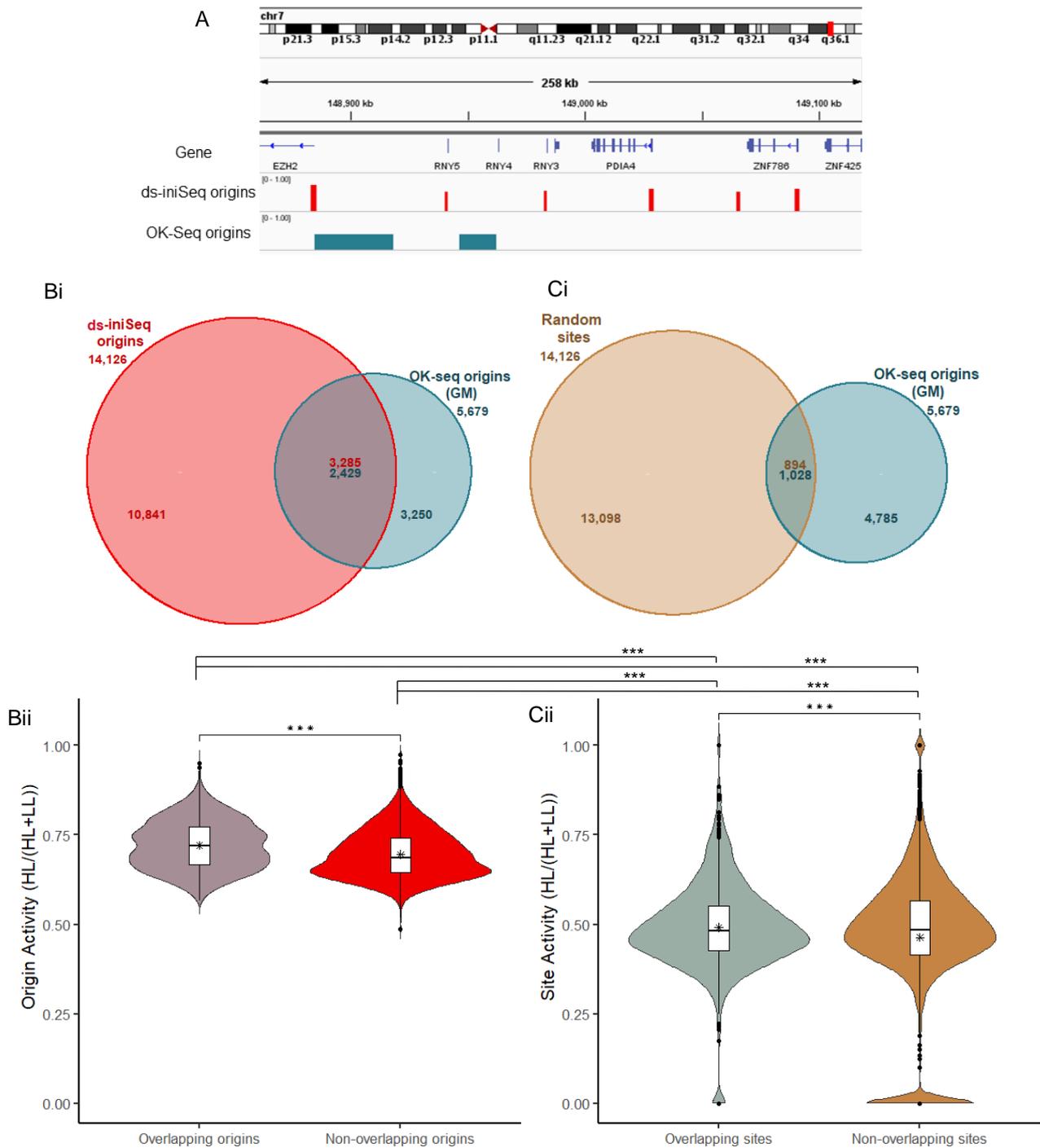
The ds-iniSeq origin relative activities of those that overlapped with SNS-seq origins were significantly higher than that of non-overlapping origins (Fig.5.3Bii). The distribution of origin activities for the non-overlapping origins was almost exclusively at the lower activity values, whereas the overlapping origins activities were more evenly distributed across the range of origin activities. This finding concurred with observations of ds-iniSeq origin activities that did and did not overlap with iniSeq origins (Fig.5.2Bii). Origins common to both methods were more active, and consequently possessed a greater probability of firing, supporting the stochastic model of origin activation (3,4).

The assessment of the ds-iniSeq RS activities (Fig.5.3Cii) showed that the relative activities of sites that overlapped with SNS-seq origins were significantly higher than those ds-iniSeq RS that did not overlap with SNS-seq origins. The explanation for the higher relative activities of ds-iniSeq RS colocalised with SNS-seq origins is the same as for the iniSeq overlap analysis. The relative activities of the ds-iniSeq origins that did and did not overlap with SNS-

seq origins were significantly higher than the ds-iniSeq RS that both overlapped and did not overlap with SNS-seq origins.

#### *5.2.2c OK-seq overlap*

I compared the overlap of ds-iniSeq origins and RS with origins identified OK-seq (OK-seq origins) from the GM06990 cell line, to assess their concordance (Fig.5.4).



Ds-iniSeq origins and OK-seq origins were visualised on the human genome in IGV (Fig.5.4A), where there were far fewer OK-seq origins (2) than ds-iniSeq origins (6), however these OK-seq origins were considerably wider. This was expected as OK-seq has been documented as generating broad origin zones of up to 150Kb (7).

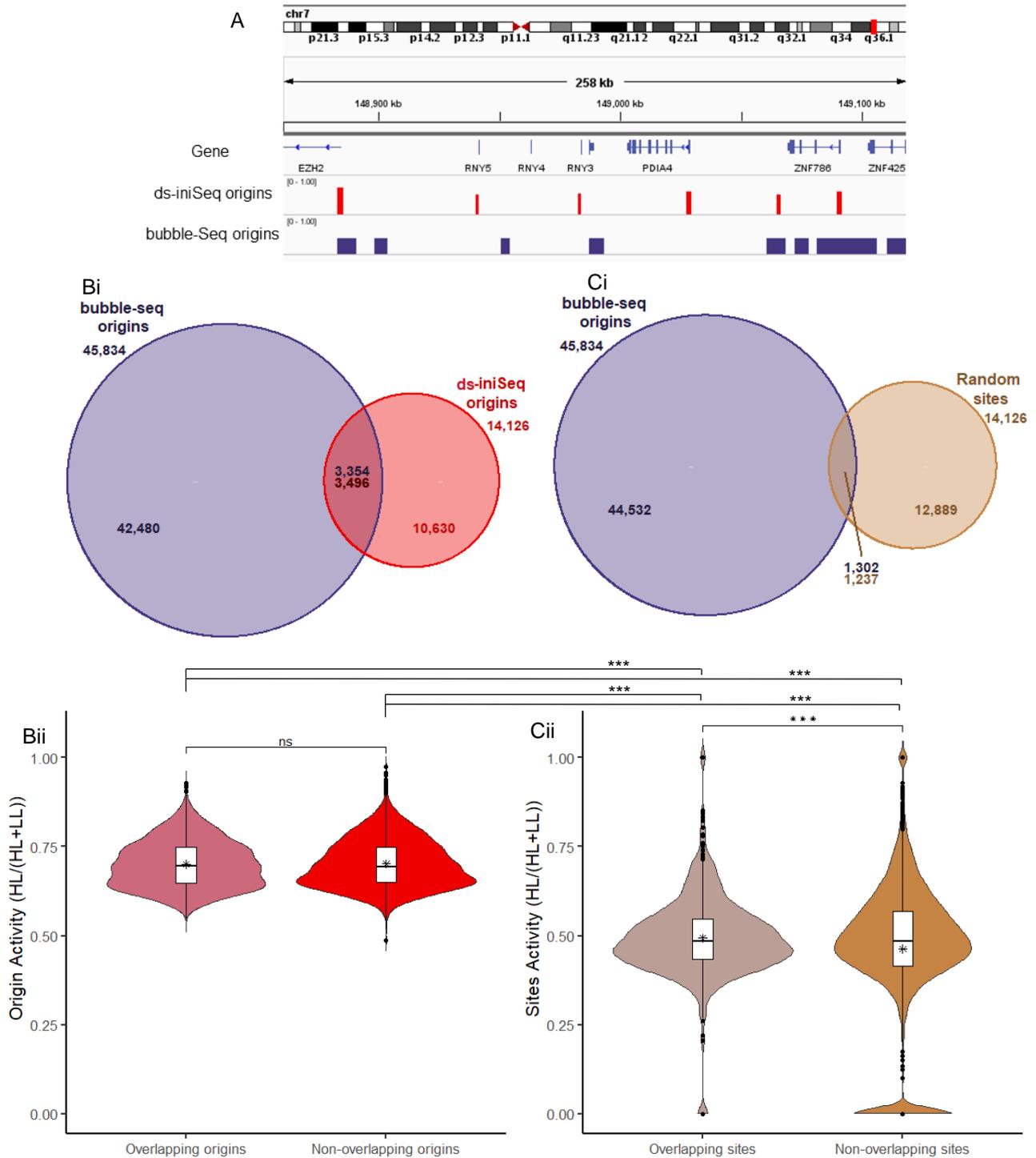
Approximately 23% of ds-iniSeq origins overlapped with ~43% of the 5,679 broad OK-seq origins (Fig.5.4Bi), whereas ~6% ds-iniSeq RS sites overlapped with 18% OK-seq origins (Fig.5.4Ci). This was a poor concordance but demonstrated a modest enrichment of ds-iniSeq origins with the OK-seq origins; ~17% of ds-iniSeq origins specifically overlapped with SNS-seq origins and did not do so by chance. This was consistent with the overlap of the iniSeq origins with the same OK-seq origin sites (1) and was directly comparable as both analyses were compared to the same OK-seq data.

The ds-iniSeq origin relative activities of those that overlapped with OK-seq origins were significantly more active than the non-overlapping ds-iniSeq origins (Fig.5.4Bii). The non-overlapping origins showed an accumulation of activities at lower levels whereas the overlapping ds-iniSeq origins had more even distributed activities. Again, supporting the stochastic model for origin firing (3,4).

The assessment of the ds-iniSeq RS activities (Fig.5.4Cii) showed that the relative activities of sites that overlapped with OK-seq origins were significantly higher than those ds-iniSeq RS that did not overlap with OK-seq origins. The explanation for the higher activities of ds-iniSeq RS colocalised with OK-seq origins is the same as for the iniSeq and SNS-seq overlap analysis. The relative activities of the ds-iniSeq origins that did and did not overlap with OK-seq origins were significantly higher than the ds-iniSeq RS that both overlapped and did not overlap with OK-seq origins.

#### *5.2.2d Bubble-seq overlap*

An overlap analysis of ds-iniSeq origins and RS with origins identified by bubble-seq (bubble-seq origins) conducted on GM06990 cells is shown in Fig.5.5.



**Figure 5.5:** The called origins found in replicate 1 that overlap with the called origins in replicates 2 and 3, were compared to the combined (and unique) origins found in the three bubble-Seq replicates (9). (A) An IGV image of the ds-iniSeq (red) and bubble-Seq (purple) origins at the region near the EZH2 gene. (Bi) The number of ds-iniSeq origins that overlap with bubble-Seq origins (red) and vice versa (purple). (Bii) The origin activity of the ds-iniSeq origins that do (overlapping origins) and do not (non-overlapping origins) overlap with bubble-Seq origins. The mean is indicated with an \* and a student's T-test was conducted; \*\*\* indicates  $p < 0.001$ . (Ci) The number of ds-iniSeq random sites that overlap with bubble-Seq origins (ochre) and vice versa (purple). (Cii) The origin activity of the ds-iniSeq random sites that do (overlapping sites) and do not (non-overlapping sites) overlap with bubble-Seq origins. The mean is indicated with an \* and a student's T-test was conducted; \*\*\* indicates  $p < 0.001$ . An ANOVA and subsequent Tukey's post-hoc test was performed to assess significance between ds-iniSeq origins (Bii) and their corresponding random sites (Cii); the Tukey's test results are shown on the plot; \*\*\* indicates  $p < 0.001$ .

The ds-iniSeq and bubble-seq origins were visualised on the human genome (IGV) (Fig.5.5A) and showed a similar number of origins from both methods, however the bubble-seq origins were wider. Some ds-iniSeq origins visually overlapped with the bubble-seq origins.

Genome-wide overlap analysis showed ~25% ds-iniSeq origins overlapped with ~7% of 45,834 bubble-seq origins (Fig.5.5Bi). Further overlap analysis showed that ~9% ds-iniSeq RS overlapped with ~3% bubble-seq origins (Fig.5.5Ci). With this accounted for, ~16% of ds-iniSeq origins specifically overlapped with bubble-seq origins. The concordance of ds-iniSeq origins with bubble-seq origins was poor compared to iniSeq origins with the same bubble-seq origins, and SNS-seq origins with bubble-seq origins (1,5,8).

Ds-iniSeq origin relative activities (Fig.5.5Bii) of those that did and did not overlap with bubble-seq origins showed no significant difference. However, those non-overlapping ds-iniSeq origins demonstrated a marginal accumulation at lower activities. The relative activities of ds-iniSeq RS (Fig.5.5Cii) that did and did not overlap with bubble-seq origins were significantly lower than that of the ds-iniSeq origins that did and did not overlap with bubble-seq origins.

From the data in Figs.5.2-5.5, I conclude that the highest and most specific concordance of ds-iniSeq origins was with SNS-seq, which was highest of any two origin identification methods. The next highest concordance was with iniSeq, followed by bubble- and then OK-seq. The high concordance of ds-iniSeq origins with SNS-seq and iniSeq origins increased the confidence that ds-iniSeq was truly selecting for DNA replication origins.

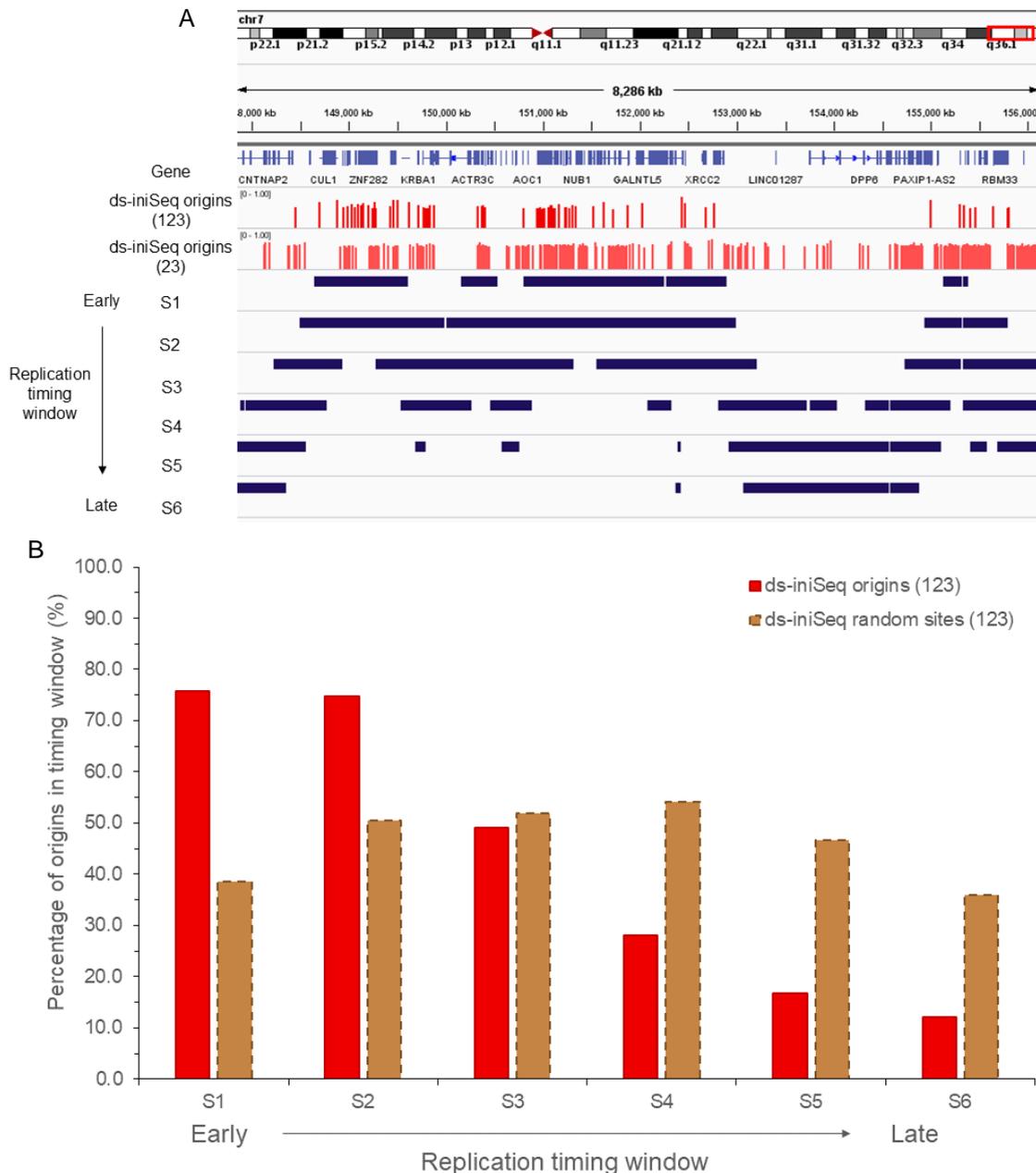
The lower concordance of the ds-iniSeq origins with OK-seq and bubble-seq origins may result, in part, from the larger widths of the OK-seq and bubble-seq origins, when compared to the narrow ds-iniSeq origins (and the iniSeq and SNS-seq origins). Additionally, the ds-iniSeq origins were most likely to select for early firing/initiating origins, whereas OK-seq and bubble-seq origins also select for late firing/initiating origins (9). However, SNS-seq was more efficient at identifying later firing/initiating origins (2,5,6) and the concordance of ds-iniSeq with SNS-seq origins was extremely high (84% after accounting for random overlap), suggesting that the timing of origin firing may not be a substantial contributory factor. However, cell lineage derived differences in origin firing may be a more likely contributor.

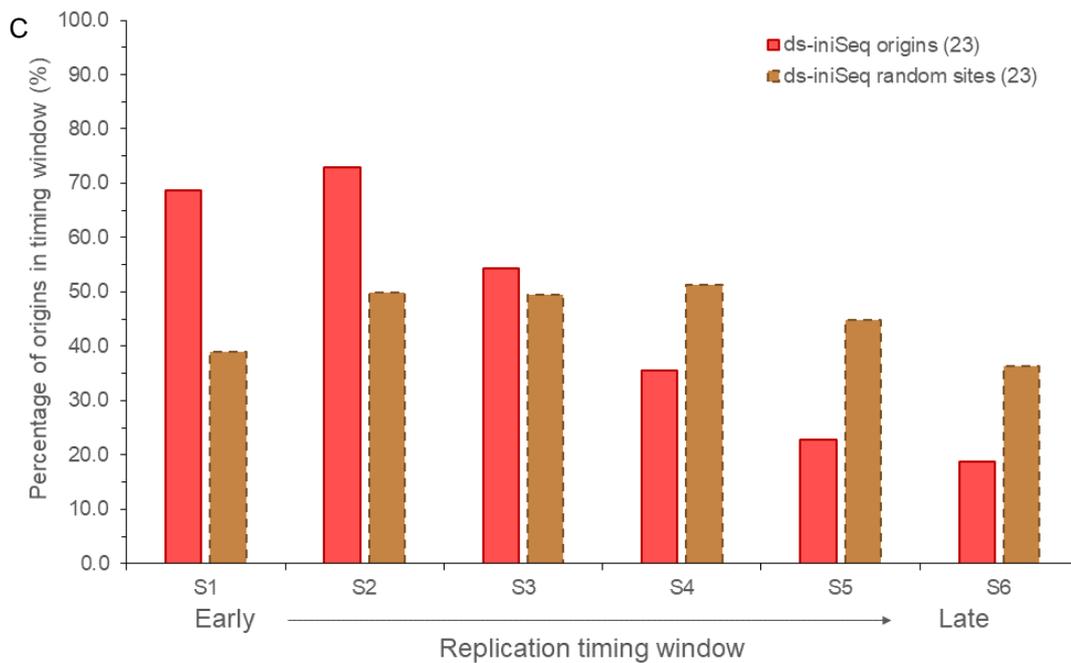
Furthermore, bubble-seq has a bias against origins in small fragments and those origins asymmetrically situated on them (9). As my ds-iniSeq origins consisted of smaller fragments, this may also explain the lower concordance and effect on origin activity of overlapping and non-overlapping ds-iniSeq origins.

### 5.2.3 Replication timing

Human genomic DNA replication takes place across 8-10 hours (10). The ds-*iniSeq* replication reaction was 15 minutes and used late G1-phase nuclei template nuclei, so it is probable that the ds-*iniSeq* origins initiated in early S-phase. In the previous chapter, I asserted that the replicates 2 and 3 represented a later timepoint in DNA replication possibly resulting from more efficient replication reactions.

I tested these hypotheses by conducting overlap analyses of ds-*iniSeq* (replicate 1 origins overlapping replicate 2 and 3 origins) and ds-*iniSeq*23 origins (replicate 2 origins overlapping replicate 3 origins only) with repli-*seq* (HeLa cell line) data (Fig.5.6) generated Dellino *et al* (11). This repli-*seq* data separated DNA replication into 6 replication timing windows, from early (S1) to late S-phase (S6).





**Figure 5.6:** The replicate 1 origins overlapping replicates 2 and 3 (ds-iniSeq origins 123) and the replicate 2 origins found in replicate 3 (and not replicate 1) (ds-iniSeq 23) were compared to replication timing profiles/windows (S1 (early replicating) to S6 (late replicating)) from HeLa cells (11). (A) These origins 123 (red) and 23 (salmon red), and the replication timing profiles (indigo) at a region on chromosome 7 were visualised on the IGV. (B) The percentage of ds-iniSeq 123 origins (red) and random sites (ds-iniSeq random 123; ochre) which were found in each replication timing window. (C) The percentage of ds-iniSeq 23 origins (salmon red) and random sites (ds-iniSeq random 23; ochre) which were found in each replication timing window.

The IGV image (Fig.5.6A) displays the visualised ds-iniSeq and ds-iniSeq23 origins and the replication timing window sites determined by repli-seq. There were more ds-iniSeq23 origins than ds-iniSeq origins, consistent with total numbers genome-wide (ds-iniSeq23 origins=56,505; ds-iniSeq origins=14,126). Ds-iniSeq origins predominantly appeared in earlier replication timing windows (S1-S3/4), whereas ds-iniSeq23 origins extended further into the later replication (S4-S6) timing windows, which is consistent with my hypotheses.

The ds-iniSeq origins were predominantly found in early replicating timing windows (S1 75.8%; S2 74.8%). Only 16.8% and 12.0% of ds-iniSeq origins were in late replication timing windows (S5&S6 respectively). The percentage of ds-iniSeq origins that were present in the mid-replication timing windows decreased rapidly as replication timing progressed (49.0% in S3 & 28.0% in S4). The ds-iniSeq RS showed very little difference in the percentage of sites found in each replication timing window (36.0%-54.2%) (Fig.5.6B), indicating there was a specific enrichment/overrepresentation of ds-iniSeq origins in the early- and to a lesser degree, mid-replication timing windows. There was a specific depletion/underrepresentation of ds-iniSeq origins in the late timing windows. This finding was almost identical to the iniSeq replication timing data (1). These observations supported my hypothesis of that the ds-iniSeq method selected/bias for earlier replicating/firing origins.

Ds-*iniSeq* origins present in later replication timing windows could consist of: genuine origins that overcame an environment associated with later origin firing/initiation; replicated sites that resulted from contaminating S-phase nuclei during the replication reaction; sites that were actually false positives; or a mixture.

It was difficult to distinguish between genuine origins and false positives at any replication timing window, but false positives were less likely as the ds-*iniSeq* origins were consistently found in all 3 replicates. S-phase contaminants could potentially have an influential effect on all ds-*iniSeq* origins. The human cell-free system, on which the ds-*iniSeq* replication reactions were based, has shown that approximately 50% of the synchronised G1 phase nuclei undergo DNA replication and that <5% of all nuclei in a replication reaction were S-phase contaminating nuclei (12–15). Therefore, there was potential that up to 5% of the replicated DNA isolated during ds-*iniSeq* could result from S-phase contaminants.

To establish the effect/influence of S-phase contaminants on these ds-*iniSeq* data, one could conduct a standard ds-*iniSeq* reaction in the absence of cytosolic extract, which would account for replication of sites resulting from S-phase contaminating nuclei only. After overcoming any potential issues with insufficient DNA yields, these sites could then be compared to the ds-*iniSeq* and ds-*iniSeq*23 origins, in order to assess the impact of S-phase contaminating sites; time constraints prevented me from conducting these experiments. I considered that the overall level of S-phase contamination was likely to be sufficiently low and producing a list of origins present in three replicates was likely to minimise the impact of S-phase contaminants.

The ds-*iniSeq*23 origins appeared to shift to later replication timing windows, when compared the ds-*iniSeq* origins (Fig.5.6C). Although the ds-*iniSeq*23 origins were predominantly found in early replication timing windows, fewer were in S1 (68.7%) and S2 (73.0%), when compared to the ds-*iniSeq* origins (75.8% & 74.8% respectively). More ds-*iniSeq*23 origins were represented in the mid-replication timing windows (54.3% & 35.5% in S3 & S4 respectively) when compared to the ds-*iniSeq* origins (49.0% & 28.0% in S3 & S4 respectively). More ds-*iniSeq*23 origins were found in the late replication timing windows (22.9% & 18.7% in S5 & S6 respectively) when compared to the ds-*iniSeq* origins (16.8% & 12.0%). The ds-*iniSeq* RS23 (random sites corresponding to the ds-*iniSeq*23 origins) were almost identical to the ds-*iniSeq* RS with ~50% of ds-*iniSeq* RS23 present in each replication timing window, demonstrating that there was also a specific enrichment of ds-*iniSeq*23 origins in early- to mid-replication timing windows, when compared to late timing windows. However, this difference was not as pronounced as the ds-*iniSeq* origins.

These data also demonstrated an overrepresentation of ds-*iniSeq*23 origins in earlier timing windows and an underrepresentation of ds-*iniSeq*23 origins in later timing windows, when

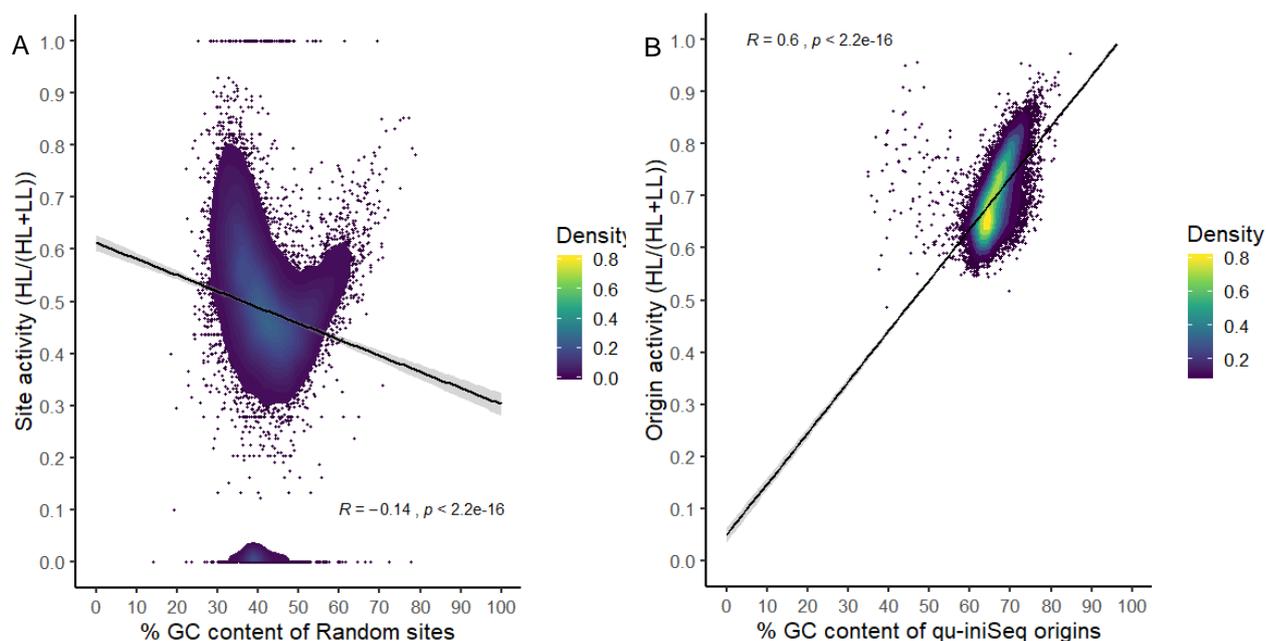
compared to the ds-iniSeq RS23. The ds-iniSeq RS23 showed the same pattern as the ds-iniSeq and iniSeq RS (1), despite the much higher number of 23 sites. This provided me with greater confidence that the RS were genuinely random.

These data support my proposal that the ds-iniSeq23 represented origins that fired in later replication timing windows; replicates 2 and 3 were effectively at a later time point in DNA replication. Those origins present in later replication timing windows could also have been genuine origins, replicated sites from contaminating S-phase nuclei or false positives. The human cell-free system displays substantial biological variability in replication reaction efficiencies between experiments that are conducted on the same batches of nuclei, which would control for S phase contamination levels and makes the argument for more efficient replication reactions in replicates 2 and 3. However, very early S-phase contaminating nuclei cannot be controlled for by confocal microscopy (observing incorporation of a fluorescent tag into newly synthesised DNA), which our lab uses to assess the suitability of nuclei for use in experimentation. I considered that the replication timing results of ds-iniSeq23 origins were most likely to be a combination of these possibilities.

Taken together, this has confirmed my hypothesis that replicates 2 and 3 represented a later timepoint in S-phase.

#### *5.2.4 Percentage GC content*

Human replication origins often have a high GC content (16). I determined the %GC content of the ds-iniSeq origins and their corresponding ds-iniSeq RS to ascertain whether %GC content affected origin selection and/or activity (Fig.5.7).



**Figure 5.7:** The % GC content of the ds-iniSeq (A) random sites and (B) origins was plotted against the random site and origin activity, respectively. A linear regression, Pearson test for correlation and ANOVA test for significance were conducted and are indicated on each plot.

The ds-iniSeq RS showed a very weak and highly significant negative correlation (-0.14) between their relative site activities and corresponding %GC content (Fig.5.7A). The mean %GC content was 41.8%, which was consistent with the %GC content of the whole human genome of 40.9% (17).

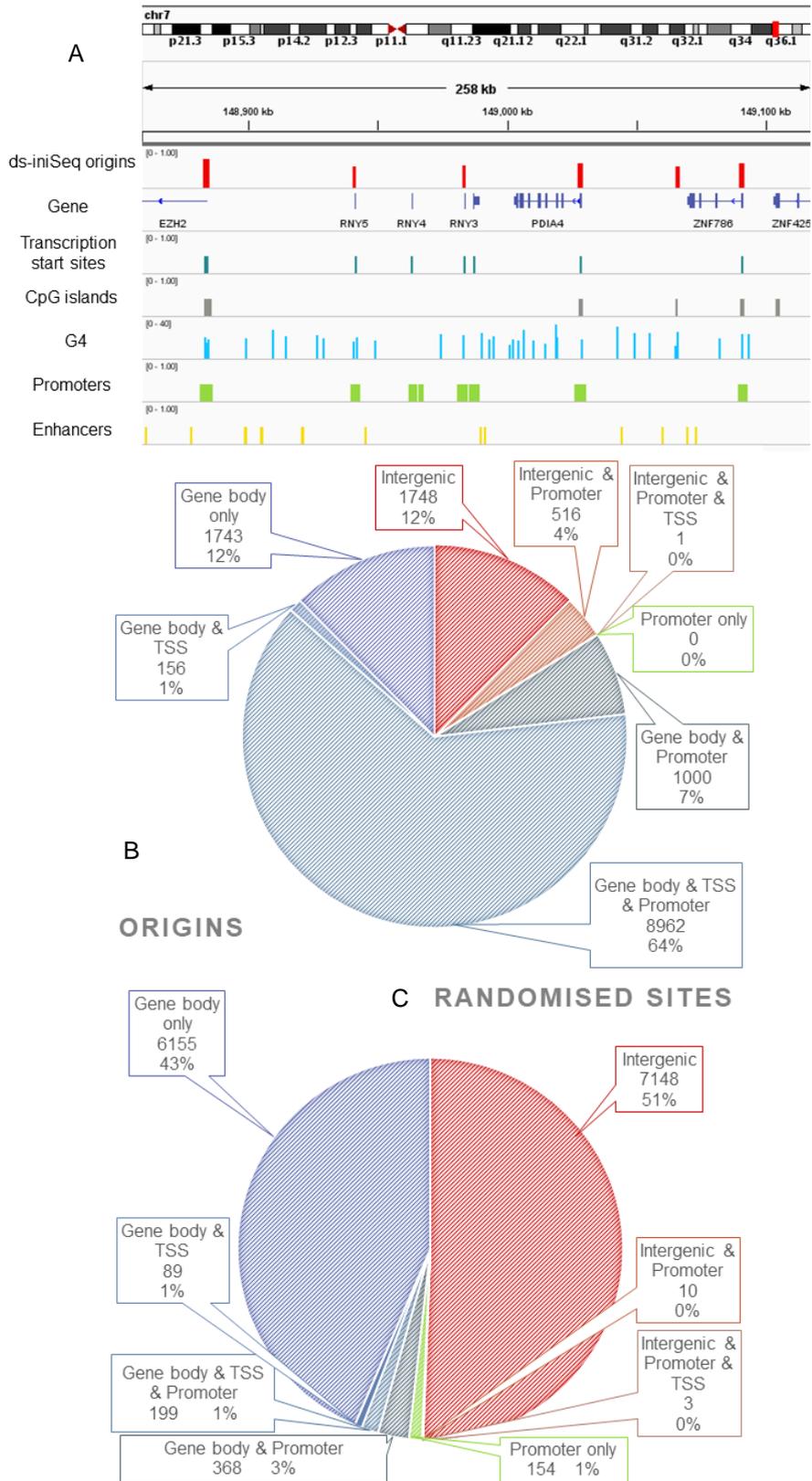
By contrast, the ds-iniSeq origins showed a strong and highly significant correlation (0.6) between their relative activities and corresponding %GC content (Fig.5.7B). The mean %GC content of the ds-iniSeq origins (66.57%), was enriched when compared to the RS (41.8%) and to that of the whole genome (40.9%) (17). This finding was consistent with the established bias of origins for GC-rich DNA. Origins have also been associated with CGIs and G4s, which as the names suggest, possess a high %GC content (>50%) (16,18,19). Additionally, the high %GC content of the origins was consistent with previous SNS-seq findings, where origins were also GC rich (20). Unlike SNS-seq, ds-iniSeq was not subject to the GC bias of lambda exonuclease digestion. However, I have yet to determine the extent, if any, of contaminating GC rich LL DNA in my HL samples.

A feature that had not previously been established was the impact of GC content had on origin activity. The relative activities of ds-iniSeq origins were more densely distributed at a narrow range of higher values than the corresponding ds-iniSeq RS. It was clear from these data, that the higher the GC content, the greater the relative origin activity of these ds-iniSeq origins; a causal relationship has yet to be established.

### *5.2.5 Comparison to genomic features*

Replication origins have previously been associated with genomic features, including genes, TSS, CGIs, G4s, promoters and enhancers (21–23). I conducted comprehensive overlap analyses of the ds-iniSeq origins with these genomic features.

The association of ds-iniSeq origins and RS with gene body, intergenic DNA, TSS and promoters is shown in Fig.5.8.



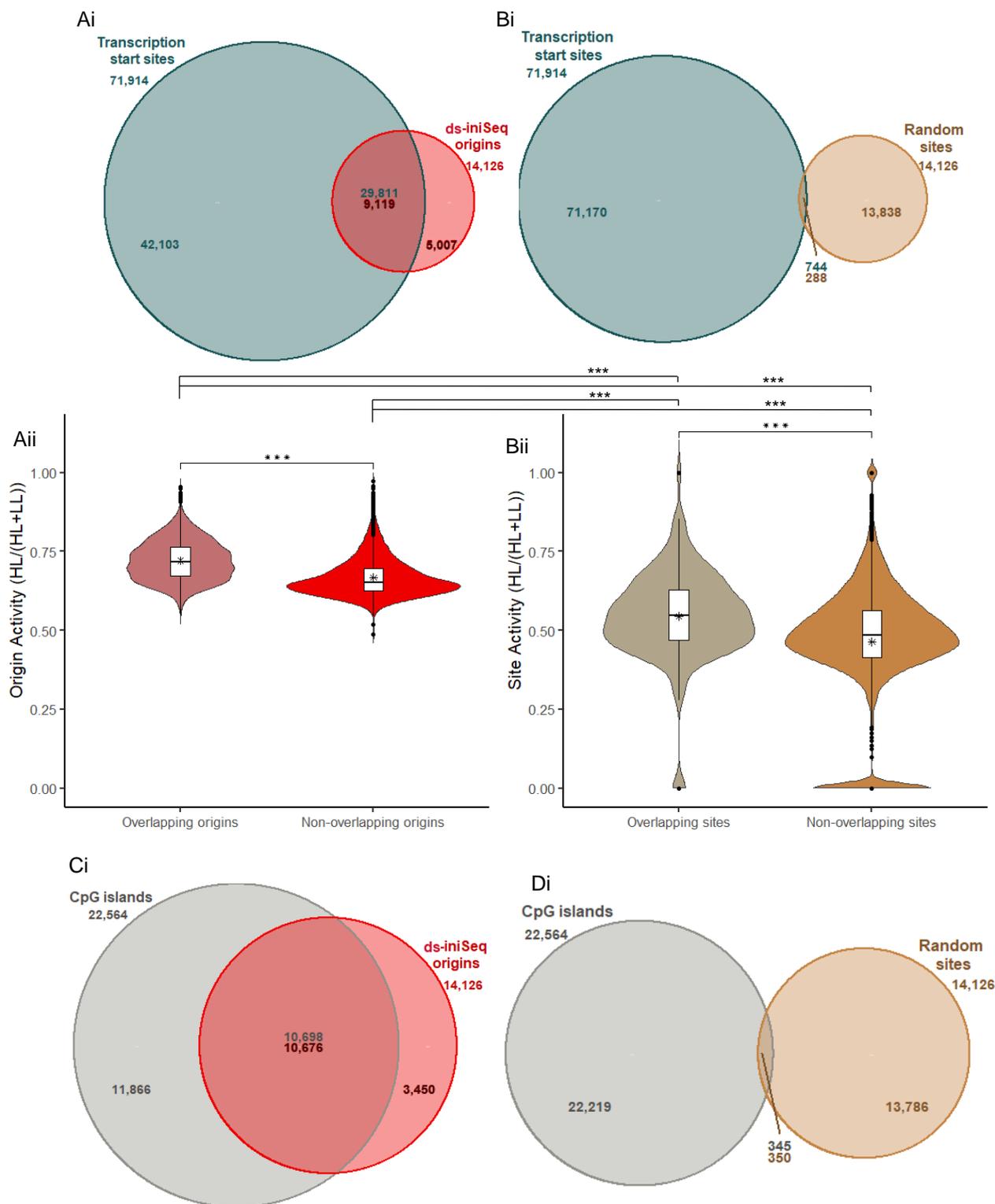
**Figure 5.8:** (A) An IGV image of the ds-iniSeq (red) and the genomic features, genes (blue-purple), transcription start sites (green-blue), CpG islands (grey), G4s (Cyan), promoters (apple green) and enhancers (gold), at the region near the EZH2 gene. (B) Pie chart showing the numbers & percentages of ds-iniSeq origins that overlap with gene bodies (blue colours) (including genes, transcription start sites and/or promoters) and intergenic DNA regions (red colours) (including intergenic DNA, transcription start sites and/or promoters). (C) Pie chart showing the numbers & percentages of ds-iniSeq random sites that overlap with gene bodies (blue colours) (including genes, transcription start sites and/or promoters) and intergenic DNA regions (red colours) (including intergenic DNA, transcription start sites and/or promoters).

The IGV image of the ds-iniSeq origins alongside genes, TSS, CGIs, predicted G4s, promoters and enhancers (Fig.5.8A) showed that ds-iniSeq origins colocalised with genes, TSS, CGIs, G4s and promoters but not enhancers.

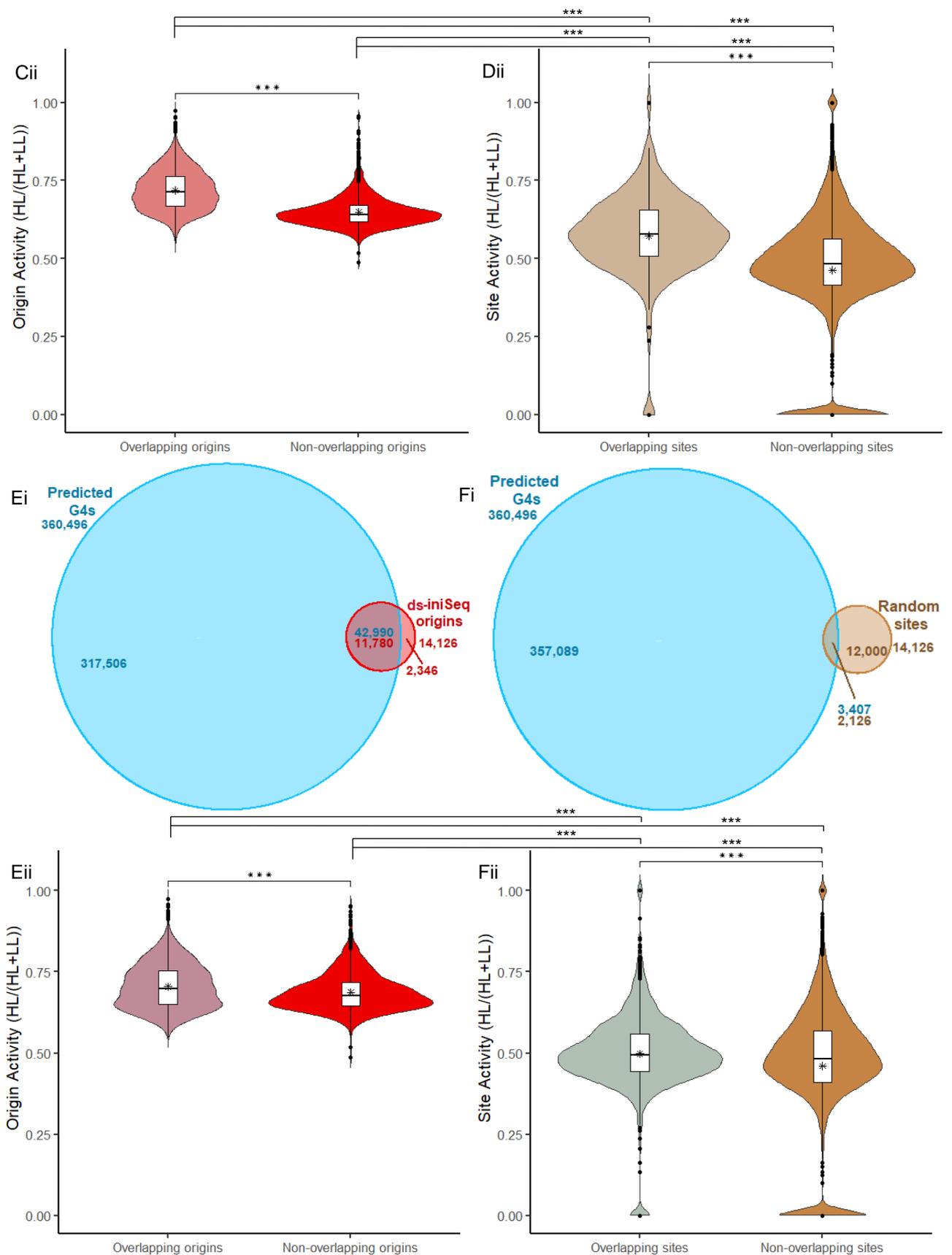
Genome-wide, approximately 84% of ds-iniSeq origins (Fig.5.8B) overlapped with gene bodies and/or TSS and/or promoters, but only 16% overlapped with intergenic DNA (and/or promoters). Comparatively, 48% of ds-iniSeq RS (Fig.5.8C) overlapped with gene bodies and/or TSS and/or promoters, whereas ~51% of the ds-iniSeq origins overlapped with intergenic DNA (and/or TSS/promoters) and the remaining 1% overlapped with only promoters.

These two analyses revealed that ds-iniSeq origins were specifically enriched at areas in the gene body and depleted in intergenic DNA. This was consistent with previous investigations of replication origins by iniSeq, that showed ~75% origins and ~54% of random sites were found in the gene body (1). Langley *et al* (1) also highlighted the strong colocalisation of iniSeq origins with TSS within the gene body (~43%), which was validated by the finding presented here; 65% ds-iniSeq origins colocalised with TSS, when compared to 2% of RS.

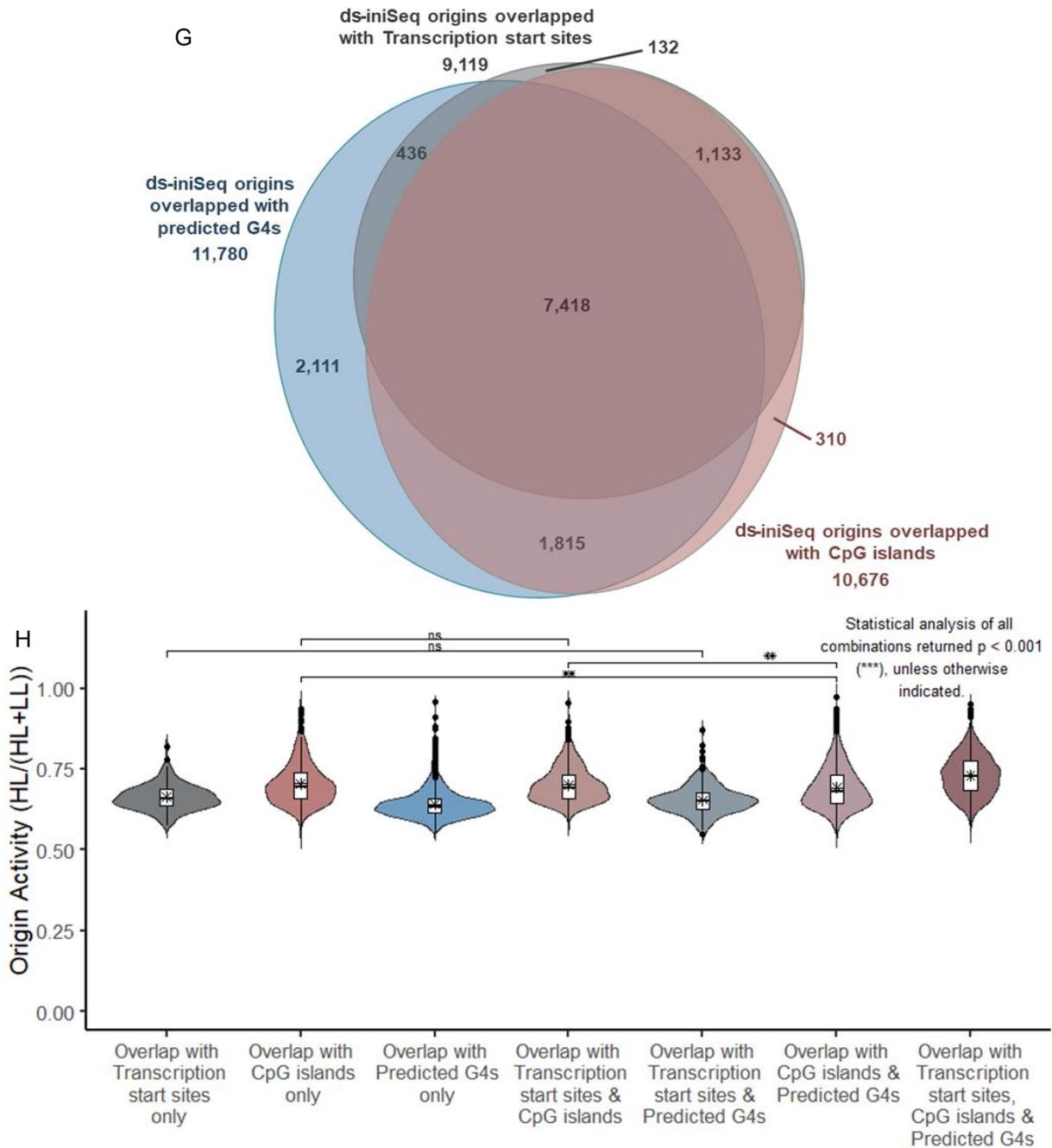
Therefore, I have performed a more in depth overlap analysis of the ds-iniSeq origins and relative activities with TSS, CGIs and G4s (Fig.5.9).



**Figure 5.9:** The called origins found in replicate 1 that overlap with the called origins in replicates 2 and 3 were compared to the combined (and unique) origins, to transcription start sites (grey green; ensemble), CpG islands (grey; Seqmonk) and predicted G4s (cyan; GG, LMB-MRC). (Ai) The number of ds-iniSeq origins that overlap with Transcription start sites (all transcripts) (red) and vice versa (grey green). (Aii) The origin activity of the ds-iniSeq origins that do (overlapping origins) and do not (non-overlapping origins) overlap with Transcription start sites. (Bi) The number of ds-iniSeq random sites that overlap with transcription start sites (ochre) and vice versa (grey green). (Bii) The origin activity of the ds-iniSeq random sites that do (overlapping sites) and do not (non-overlapping sites) overlap with transcription start sites. (Ci) The number of ds-iniSeq origins that overlap with CpG islands (red) and vice versa (grey). (Cii) The origin activity of the ds-iniSeq origins that do (overlapping origins) and do not (non-overlapping origins) overlap with CpG islands.



**Figure 5.9:** (Di) The number of ds-iniSeq random sites that overlap with CpG islands (ochre) and vice versa (grey). (Dii) The activity of the ds-iniSeq random sites that do (overlapping sites) and do not (non-overlapping sites) overlap with CpG islands. (Ei) The number of ds-iniSeq origins that overlap with G4s (red) and vice versa (cyan). (Eii) The origin activity of the ds-iniSeq origins that do (overlapping origins) and do not (non-overlapping origins) overlap with G4s. (Fi) The number of ds-iniSeq random sites that overlap with G4s (ochre) and vice versa (cyan). (Fii) The origin activity of the ds-iniSeq random sites that do (overlapping sites) and do not (non-overlapping sites) overlap with G4s.



**Figure 5.9:** (G) The 3 way overlap of origins found at transcriptions start sites, origins found at CpG islands and origins found at G4s. (H) The origin activity of origins that: only overlap transcription start sites, CpG islands or G4s; overlap transcription start sites and CpG islands, transcription start sites and G4s or overlap CpG islands and G4s; and origins that overlap all 3 genomic features. The means are indicated with an \*. An ANOVA and subsequent Tukey's post-hoc test were performed to assess significance; the Tukey's test results are shown on the plot and all analytical combinations were significant with a p values  $< 0.001$  unless indicated otherwise. \*\* indicates  $p < 0.01$  and ns indicates a not significant result. (All (ii)) The means are indicated with an \* and a student's T-test was conducted; \*\*\* indicates  $p < 0.001$ . An ANOVA and subsequent Tukey's post-hoc test was performed to assess significance between ds-iniSeq origins and their corresponding random sites; the Tukey's test results are shown on the plot; \*\*\* indicates  $p < 0.001$ .

### *5.2.5a Transcription start sites (TSS)*

Overlap analysis of ds-iniSeq origins with TSS (NB TSS were of all transcript isoforms) showed that 65% origins overlapped with 42% of the 71,914 TSS (Fig.5.9Ai). Only 2% of ds-iniSeq RS overlapped with 1% of TSS (Fig.5.9Bi) which demonstrated enrichment of ds-iniSeq origins at TSS over and above the colocalisation by random chance and consistent with previous findings (1,23).

The ds-iniSeq origin activities of those colocalised with TSS were significantly higher than those that did not overlap with TSS (Fig.5.9Aii). The overlapping origins' activities were more evenly spread distributed and higher than the non-overlapping origins, which possessed a large accumulation of origins with lower activities.

The ds-iniSeq RS (Fig.5.9Bii) of those overlapping TSS were significantly higher than the non-overlapping ds-iniSeq RS. As shown above, TSS specifically colocalised with origins possessing significantly higher activities and any random site that happened to overlap a TSS would, consequently, be more likely to possess higher relative origin activities. This explanation is applicable to the differences in relative activities of ds-iniSeq RS and any genomic feature. The relative activities of the overlapping and non-overlapping ds-iniSeq origins were both significantly higher than the relative activities of the ds-iniSeq RS that did and did not overlap TSS.

The ds-iniSeq origin relative activities of the overlapping and non-overlapping origins showed that their colocalisation with TSS was associated with increased relative origin activities. From the determination of ds-iniSeq origins with origins identified by alternative NGS methods (Fig.5.2-5.5), those with higher relative activities fired more consistently and were more efficient. Using relative origin activity as a proxy for probability of origin firing, suggests that the ds-iniSeq origins colocalised with TSS had a higher probability of firing. This could indicate that association with TSS is a character responsible for origin specification and/or firing/initiation.

### *5.2.5b CpG islands (CGIs)*

Around 76% of ds-iniSeq origins (Fig.5.9Ci) overlapped with 47% of the 22,564 CGIs, while only 2.5% of ds-iniSeq RS overlapped with 1.5% of CGIs (Fig.5.9Di). As with TSS, there was a specific enrichment of ds-iniSeq origins at CGIs. This was consistent with previous findings, where CGIs were associated with >50% origins; specifically early firing origins (18,24,25).

The ds-iniSeq origin relative activities that overlapped with CGIs were significantly higher than those that did not (Fig.5.9Cii). The distribution of the overlapping origins' activities was very similar to that of the TSS overlapping origins. Those ds-iniSeq origins not overlapping

CGIs displayed a very large accumulation of low origin activities with a small spread. The distribution and significant difference of these CGI-overlapping and non-overlapping ds-iniSeq RS (Fig.5.9Dii) were almost identical to those sites that did and did not overlap with TSS (Fig.5.9Bii). The relative activities of these overlapping and non-overlapping ds-iniSeq RS were significantly lower than the ds-iniSeq origins that did and did not overlap CGIs.

The overlapping origins showed that the colocalisation of CGIs resulted in greater relative origin activity and therefore, more efficient, thus implying that the presence of CGIs at ds-iniSeq origins resulted in a greater probability of firing. This agreed with prior research that showed CGIs were regularly found at more efficient origins (23). However, CGIs have previously been shown to colocalise with 60-70% of vertebrate gene promoters (26). Therefore, it is hard to distinguish between the effect of CGIs, on replication origins, from that of RNA transcription (26) and this must be investigated further.

#### 5.2.5c G-quadruplexes (G4s)

Roughly 84% of ds-iniSeq origins overlapped with 12% of the 360,496 predicted G4s (Fig.5.9Ei) but 15% ds-iniSeq RS overlapped with 0.9% of G4s (Fig.5.9Fi), which showed that ds-iniSeq origins were enriched at G4 motifs. This concurred with current understanding; Langley *et al* (1) and Valton & Prioleau (19) found that 48% and 80% (respectively) of origins colocalised with G4s.

The relative activities of ds-iniSeq origins overlapping predicted G4s were significantly higher than those that did not (Fig.5.9Eii) with a more evenly spread distribution of relative activities. The non-overlapping origins possessed a large accumulation of lower relative activity origins. The distribution and significant difference of G4-overlapping and non-overlapping ds-iniSeq RS (Fig.5.9Fii) were almost identical to those sites that did and did not overlap with TSS (Fig.5.9Bii). The relative activities of these overlapping and non-overlapping ds-iniSeq RS were significantly lower than the ds-iniSeq origins that did and did not overlap G4s.

Those ds-iniSeq origins overlapping G4s possessed greater origin activities than the non-overlapping origins. Again, this implies that the presence of G4 at ds-iniSeq origins is associated with a greater probability of firing, suggesting that G4s played a role on origin selection and/or efficiency/activity. Although, the difference was not as large as those observed with TSS and CGIs, which may indicate that association with G4s may not be as influential on origin firing as association with TSS or CGIs.

There have been >370,000 predicted G4 motifs identified in the human genome (27). This vast excess of predicted G4s over origins suggests that they do not constitute the sole feature responsible for origin specification/activation.

#### 5.2.5d Multiple overlap analyses

The TSS, CGI and G4 overlap analysis (Fig.5.9A-F) and established research suggest that these features are influential in human replication origin specification and/or activation. I therefore assessed the overlap of origins colocalised with 1 or more of these features and for the first time, evaluated which feature(s) had the greatest impact on origin activity.

Overlap analysis of TSS-overlapping (TSS-origins), CGI-overlapping (CGI-origins) and G4-overlapping ds-iniSeq origins (G4-origins) (Fig.5.9Gi) showed that most origins colocalised with all three features (7,148). The CGI-origins showed a moderate overlap with TSS-origins (1,133) or G4-origins (1,815). Whereas the TSS-origins showed a small overlap with G4-origins only (436).

I assessed the most significant/influential feature on ds-iniSeq origin activation by comparing the relative activities of TSS-origins, CGI-origins and G4-origins overlaps in the different overlap conditions (Fig.5.9Gii).

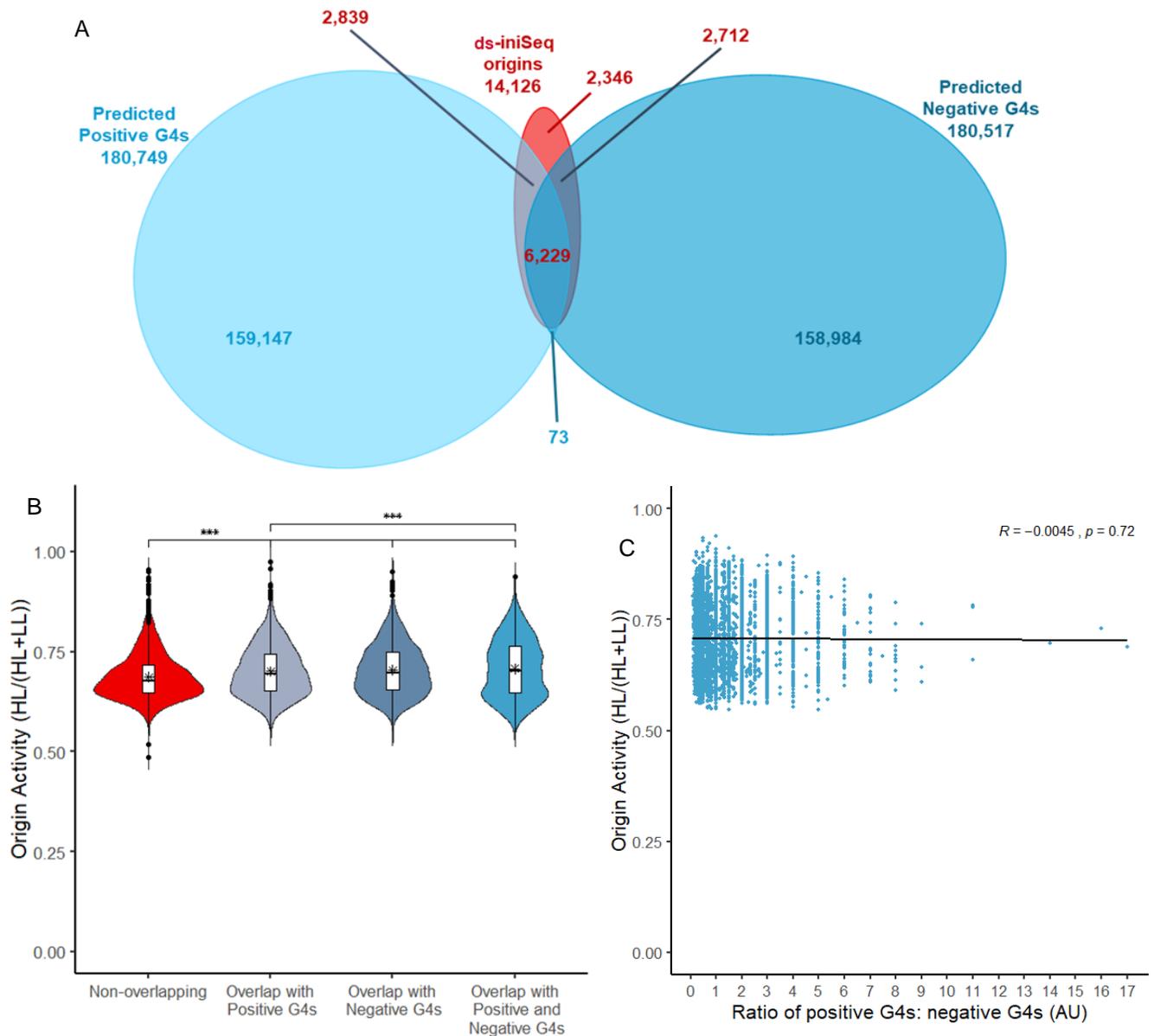
Origins that colocalised with TSS and/or G4s only displayed a similar distribution of origin activities, with an accumulation of lower activity origins and were highly significantly lower ( $p < 0.001$ ) than that of all other origins. The ds-iniSeq origins activities overlapping CGIs only and CGIs+TSS were not significantly different from one another and were significantly higher than the origins overlapping TSS and/or G4s, but significantly less active than origins that overlapped CGIs, TSS and G4s. The ds-iniSeq origin activities of those that overlapped TSS, CGIs and G4s possessed the significantly highest activities of all overlap groups. Together they show that the presence of CGIs is associated with increased origin activity compared to TSS and G4s, and the origins associated with all three features together are the most active.

These data (Fig.5.9G) clearly showed that the ds-iniSeq origins colocalised with CGIs and G4s possessed the highest relative activities, suggesting that these three features are crucial in replication origin specification and activation. It is possible that the combination of TSS, CGIs and G4s create an environment that is most conducive for efficient origin firing/initiation. Moreover, the lack of a primary consensus sequence for origin specification in humans implies that genomic features are not the characteristics that drive replication origin specification and activation; epigenetics have been strongly implicated (27,28). In particular CGIs strongly colocalise with the early replication origin firing-associated and active transcription mark, H3K4me3 and the late replication origin firing-associated and transcriptional silencing mark H3K27me3 (29–31), resulting in a bivalent chromatin state (32–34). This suggests a complex interaction between CGIs and known transcription- and DNA replication-associated epigenetic marks.

It was apparent from the data that CGIs were the dominant feature in influencing relative origin activity; all CGI-overlapping origins were more active/efficient than those without. This adds further weight to the findings presented in Sequiera-Mendes *et al* (23) which identified CGIs as a key player for origin activity and to Cadoret *et al* (35) who showed that origin density strongly correlated with GC-rich DNA clusters. Although this identified CGIs (Fig.5.9) as the front runner in defining origin features, it did not distinguish between the effect of CGIs from that of promoters and transcription; CGIs colocalise with 60-70% of promoters (26). Delgado *et al* (36) found that bulk CGIs co-ordinately replicated during early S-phase and proposed CGIs as initiation sites for both transcription and DNA replication. Further work should be conducted to investigate the distinction between those that orchestrate origin firing and those that regulate transcription.

#### 5.2.6 Polarity of G4s

Predicted G4 sites have been associated with replication origins in previous analyses (1). The ds-iniSeq origins were highly colocalised with the predicted G4s. These G4s were either located on the positive sense or negative antisense DNA strand. It is actively discussed that G4 orientation/strandedness may influence replication origins (21,37). I performed overlap analyses to determine whether G4 strandedness correlated with origin location and relative activity (Fig.5.10).



**Figure 5.10:** (A) The 3 way overlap of predicted G4s on the positive (light cyan) and negative DNA strands (dark cyan) with the ds-iniSeq origins (red). Values indicated on the plot in red are the number of origins that did/didn't overlap with positive and/or negative G4s and the values in light cyan are the number of positive G4s that overlap with negative G4s and didn't overlap either origins or negative G4s. The values in dark cyan indicate the number of negative G4s that did not overlap origins or positive G4s. (B) The origin activity of the ds-iniSeq origins that; did not overlap with G4s of either polarity ("non-overlapping), overlapped with only positive G4s, overlapped with only negative G4s and that overlapped with both positive and negative G4s. The means are indicated with an \*. An ANOVA and subsequent Tukey's post-hoc test were performed to assess significance; the Tukey's test results are shown on the plot and \*\*\* indicates  $p < 0.001$ . (C) The ratio of positive to negative G4s found at each overlapping origin was plotted against the corresponding origin activity. A linear regression, Pearson test for correlation and ANOVA test for significance were conducted and are indicated on each plot.

Overlap analysis of ds-iniSeq origins with positive and negative strand G4s (Fig.5.10A) showed that most origins (6,229) overlapped with positive and negative G4s. Roughly equal numbers of ds-iniSeq origins overlap with either positive (2,839) or negative (2,712) G4s.

The distribution of origin activities of ds-iniSeq origins that overlapped with either positive or negative G4s were not significantly different (Fig.5.10B). Those ds-iniSeq origins that overlapped with positive and negative G4s were the most active and had two regions of activity accumulation at lower and higher activities. The overall relative activity of these origins was not significantly higher than those origins overlapping only negative G4s but was significantly higher than those origins overlapping only positive G4s.

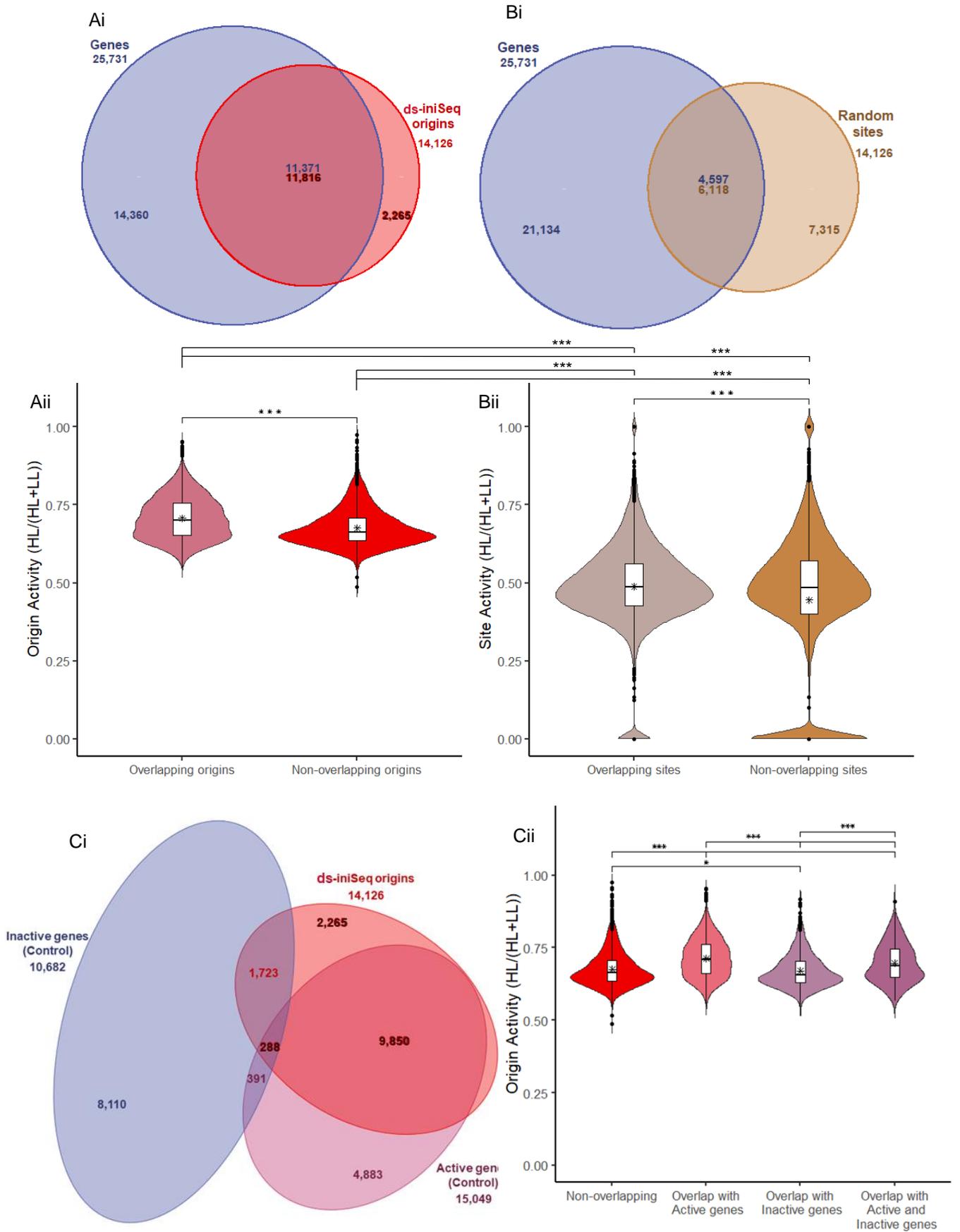
As expected, there was a substantial difference in activities between origins that did not overlap with positive and/or negative G4s when compared to those that did.

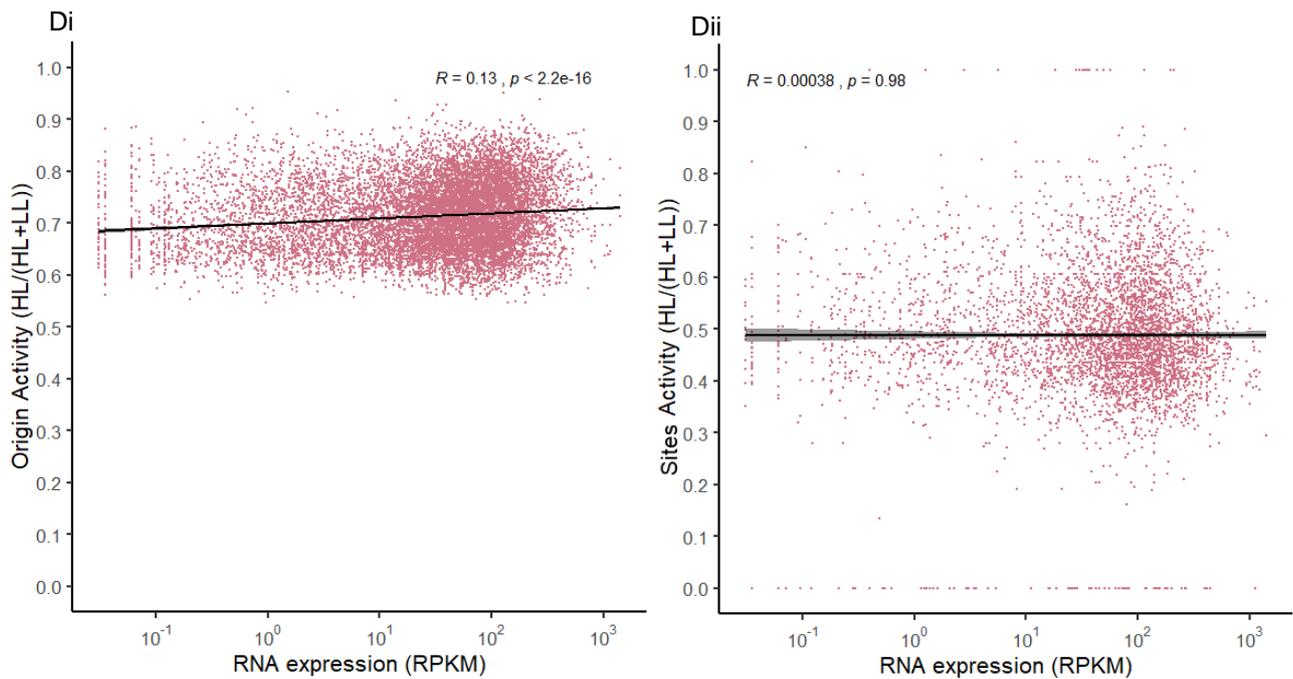
Finally, I assessed the effect the ratio of positive to negative G4s (the number of positive G4s divided by the number of negative G4s present at each origin overlapping both features) had on ds-iniSeq origin activity (Fig.5.10C) which showed no correlation between positive to negative G4 ratios and ds-iniSeq origin activity.

I conclude that the strandedness of G4s at ds-iniSeq origins did not correlate with their activity (Fig.5.10); merely the colocalisation of ds-iniSeq origins with G4s was sufficient to increase origin activity. This was a reasonable finding, as DNA replication is bi-directional process and needs access to both positive and negative DNA strands.

### 5.2.7 Genes

In previous literature and data shown earlier (Fig.5.8), replication origins have been found to colocalise with genes/gene bodies. I have carried out more in-depth analysis of the relationship between ds-iniSeq origins (and RS for comparison) and genes and their gene activities (Fig.5.11). Gene activity was determined via total RNA-seq analysis on untreated (control) EJ30 cells, in triplicate (see methods 3.13/14). The activities were quantitated using SeqMonk's RNA quantitation pipeline and genes were classed as inactive when the average number of reads per kilobase of transcript per million reads of the library (RPKM) (of 3 replicates) were equal to 0.





**Figure 5.11:** The called origins found in replicate 1 that overlap with the called origins in replicates 2 and 3 were compared to the combined (and unique) were compared to annotated genes (blue purple) where transcript isoforms were merged (ensemble). (Ai) The number of ds-iniSeq origins that overlap with annotated genes (merged transcripts) (red) and vice versa (blue purple). (Aii) The origin activity of the ds-iniSeq origins that do (overlapping origins) and do not (non-overlapping origins) overlap with annotated genes (merged transcripts). The mean is indicated with an \* and a student's T-test was conducted; \*\*\* indicates  $p < 0.001$ . (Bi) The number of ds-iniSeq random sites that overlap with annotated genes (merged transcripts) (ochre) and vice versa (blue purple). (Bii) The site activity of the ds-iniSeq random sites that do (overlapping origins) and do not (non-overlapping origins) overlap with annotated genes (merged transcripts). The mean is indicated with an \* and a student's T-test was conducted; \*\*\* indicates  $p < 0.001$ . An ANOVA and subsequent Tukey's post-hoc test was performed to assess significance between ds-iniSeq origins (Aii) and their corresponding random sites (Bii); the Tukey's test results are shown on the plot; \*\*\* indicates  $p < 0.001$ . (C) Control RNA expression was determined from the quantification (SeqMonk RNA-seq quantitation pipeline) of total RNA-seq of untreated EJ30 cells. Active (Reads per kilobase of transcript per million reads (RPKM)  $> 0$ ) and inactive (RPKM = 0) genes were determined from the control RNA expression. (i) The 3 way overlap of active genes (mauve) and inactive (blue purple) with the ds-iniSeq origins (red). (ii) The origin activity of the ds-iniSeq origins that; did not overlap with active or inactive genes ("non-overlapping), overlapped with only active genes, overlapped with only inactive genes and that overlapped with both active and inactive genes. The means are indicated with an \*. An ANOVA and subsequent Tukey's post-hoc test were performed to assess significance; the Tukey's test results are shown on the plot and \*\*\* indicates  $p < 0.001$  and \* indicates  $p < 0.05$ . (Di) The origin activity of ds-iniSeq origins overlapping genes was plotted against the control RNA expression of the corresponding gene. (Dii) The origin activity of ds-iniSeq random sites overlapping genes was plotted against the control RNA expression of the corresponding gene. (D) A linear regression, Pearson test for correlation and ANOVA test for significance were conducted and are indicated on each plot.

Most ds-iniSeq origins (84%) overlapped with 44% of the 25,731 annotated genes (Fig.5.11Ai). Conversely, 43% of ds-iniSeq RS overlapped with 33% of the annotated genes (Fig.5.11Bi). This demonstrated an enrichment of origins at genes, which is consistent with current literature (23,38).

The relative activities of the ds-iniSeq origins that overlapped with the annotated genes were significantly greater and more evenly and largely distributed than the non-overlapping origins, which possessed a large accumulation of origins with lower activities (Fig.5.11Aii). This difference in relative activity indicates that the colocalisation of a gene with a replication origin associates with increased probabilities of those origins firing. The ds-iniSeq origin relative activities that did and did not overlap with the annotated genes were both significantly higher than the overlapping and non-overlapping ds-iniSeq RS.

Although the association of DNA replication origins with genes has been established (23) and confirmed here, the effect of transcriptional activity on the activity/efficiency of replication origins has remained elusive. The ds-iniSeq method allowed me to address this question.

As there have been multiple findings that indicate a link between gene activity and active replication (23,28,38,39), I investigated the impact of active (RPKM >0) and inactive genes (RPKM =0) on origin specification and activation. Overlap analysis showed that 70% of ds-iniSeq origins overlapped with active genes but only 12% colocalised with inactive genes (Fig.5.11Ci), whereas 2% of ds-iniSeq origins overlapped with both active and inactive genes (these are different genes that overlap each other either on the same or opposite DNA strand).

The relative activities of ds-iniSeq origins overlapping inactive genes or not overlapping genes at all were very similar (Fig.5.11Cii). They both possessed similar activity distributions, with an accumulation of lower activities and few higher activities and were the least significantly different from one another ( $p=0.028$ ). The ds-iniSeq origins that overlapped with active genes only exhibited the significantly highest relative activity when compared to other overlap groups (non-overlapping:  $p=9*10^{-7}$ ; inactive genes:  $p<2.2*10^{-16}$ ; both:  $p=9.06*10^{-5}$ ). The distribution of activities was the most evenly spread. The relative activities of the ds-iniSeq that overlapped with both active and inactive genes possessed the next highest relative activity with an evenly spread distribution, but they did possess a slight accumulation of lower activity origins.

Fig.5.11C showed that the ds-iniSeq origins were not only enriched at transcriptionally active genes when compared to inactive genes, but they were also more active. On the rare occasion when a ds-iniSeq origin overlapped both an active and inactive gene, the effect of the active gene on relative origin activity was dominant over the inactive gene effect.

I established that higher origin activity associated with active genes (Fig.5.11C). I then investigated the effect that the extent of transcriptional activity (of the active genes) had on origin activity, to ascertain whether high transcriptional activity did or did not correlate with replication origin activity.

The relative activities of ds-iniSeq origins that overlapped with annotated genes were plotted against the corresponding RNA expression levels of each active gene transcript (NB the gene transcript isoforms were merged) (Fig.5.11Di). Despite the strong concordance between ds-iniSeq origins and genes, Fig.5.11Di showed a very weak positive correlation ( $R=0.13$ ) between relative origin activity and RNA expression (ie transcriptional/gene activity).

The ds-iniSeq RS relative activities of those overlapping annotated genes were plotted against the corresponding RNA expression levels (Fig.5.11Dii). It showed no correlation ( $R=0.00038$ ) between relative site activity and RNA expression (gene activity). Thus, showing that there was a negligible difference between the correlation of origin and transcriptional activity and that of randomised sites.

It is highly probable that transcription and replication is regulated and coordinated (23,28,38,40). Some research suggests that replication origins are prevalent at moderately active genes (23), whereas others have shown that DNA replication preferentially takes place at TSS with high RNA polymerase II occupancy (39). However, that preference diminishes at highly transcribed DNA (38). The current literature creates a moderately ambiguous picture of the impact of transcriptional activity on DNA replication.

Here (Fig.5.11D), I conclude that there was almost no correlation between RNA transcription activity at active genes overlapped with ds-iniSeq origins and corresponding relative origin activity. This clearly indicates that transcriptional activity and ds-iniSeq origin activity are independent from one another, despite them taking place on the same region of DNA; one did not negate the other.

This result also helps address the inability to distinguish between the effect of CGI (60-70% are at gene promoters (26)) and of RNA transcription on DNA replication origin specification and activity. As origin activity was independent of transcriptional activity, it has invalidated the argument that the higher relative origin activity of CGI overlapping origins could have been due to the influence of active transcription. I can now conclude that the presence of CGIs at ds-iniSeq origins solely correlated with increased relative activity/efficiency and, by proxy, increased probability of origin firing/initiation.

Taken together, Fig.5.11C/D revealed that ds-iniSeq origins strongly colocalised with active genes, which correlated with high replication activity. However, the levels of transcriptional activity of those active genes did not correlate with relative origin activity.

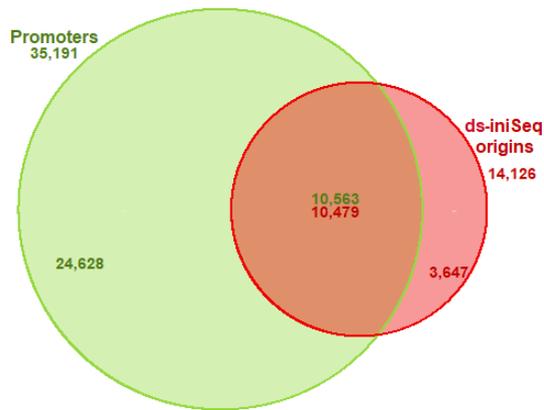
The nuclei extracted from EJ30 cells used in these ds-iniSeq experiments were treated with mimosine, to synchronise them in late G1. I repeated the comparison of ds-iniSeq origins and RS with transcriptional/gene activity (Fig.5.11C) and overlap with active and/or inactive

genes (Fig.5.11D), using the RNA expression determined by the RNA-seq of mimosine-treated cells (appendix Fig.A5.2). The RNA expression was quantified, and active and inactive genes were determined using the same methods as the untreated cells. These showed almost identical results to the untreated (control) counterpart. Consequently, I was able to conclude that the synchronisation of cells to late G1, did not affect the relationship described above between transcription and replication.

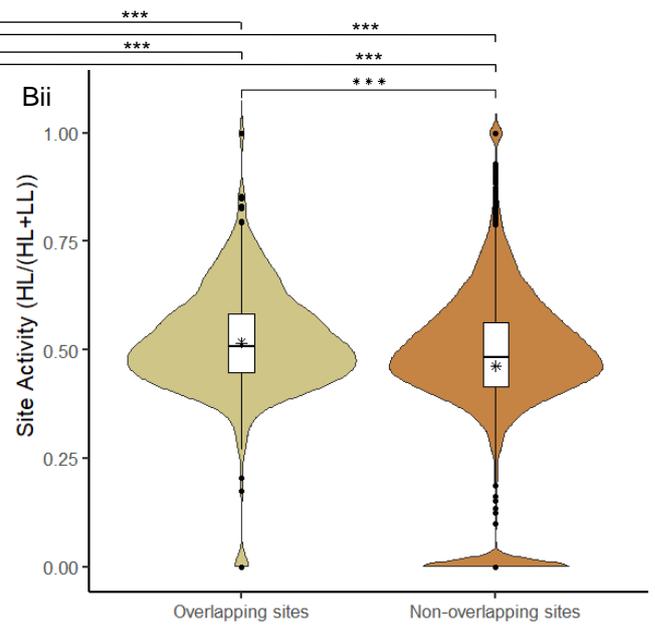
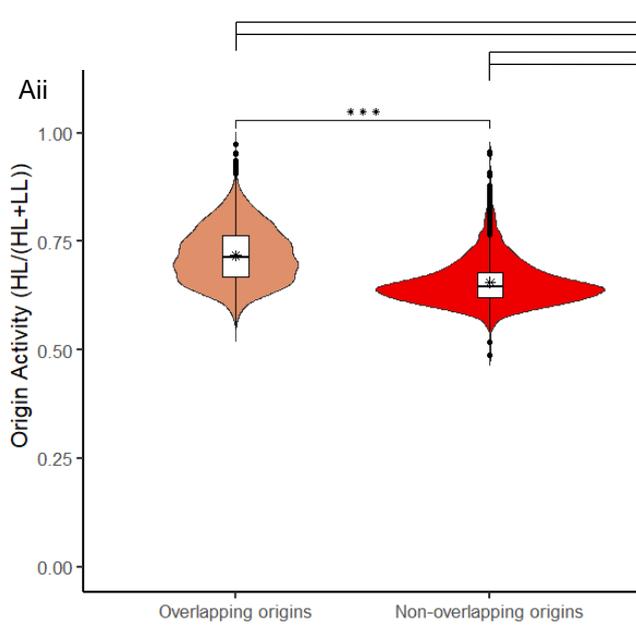
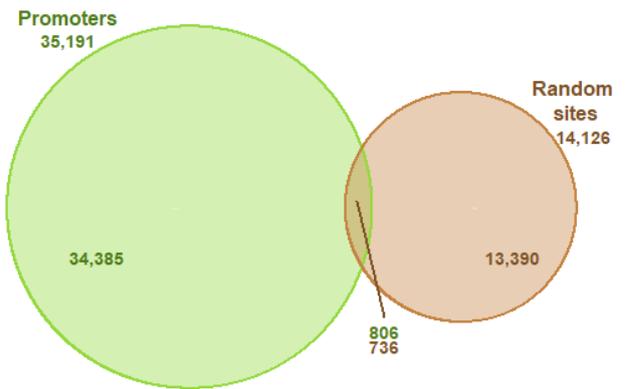
#### *5.2.8 Promoters and enhancers*

I performed an overlap analysis of ds-iniSeq origins with gene promoters and enhancers (Fig.5.12).

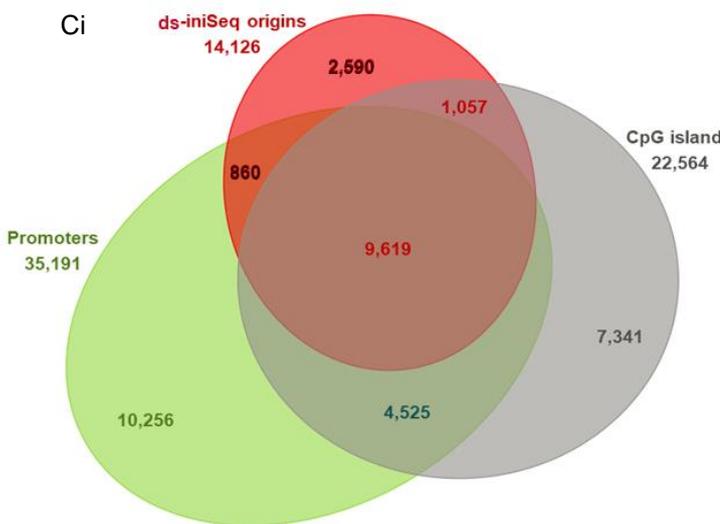
Ai



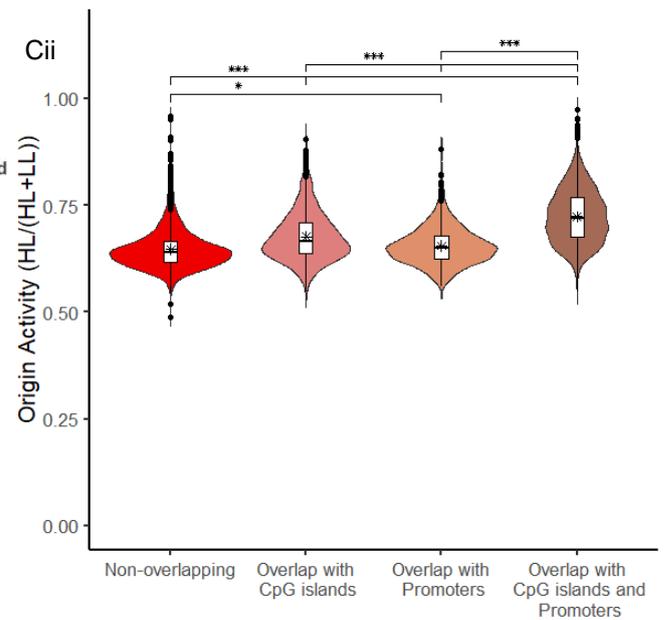
Bi

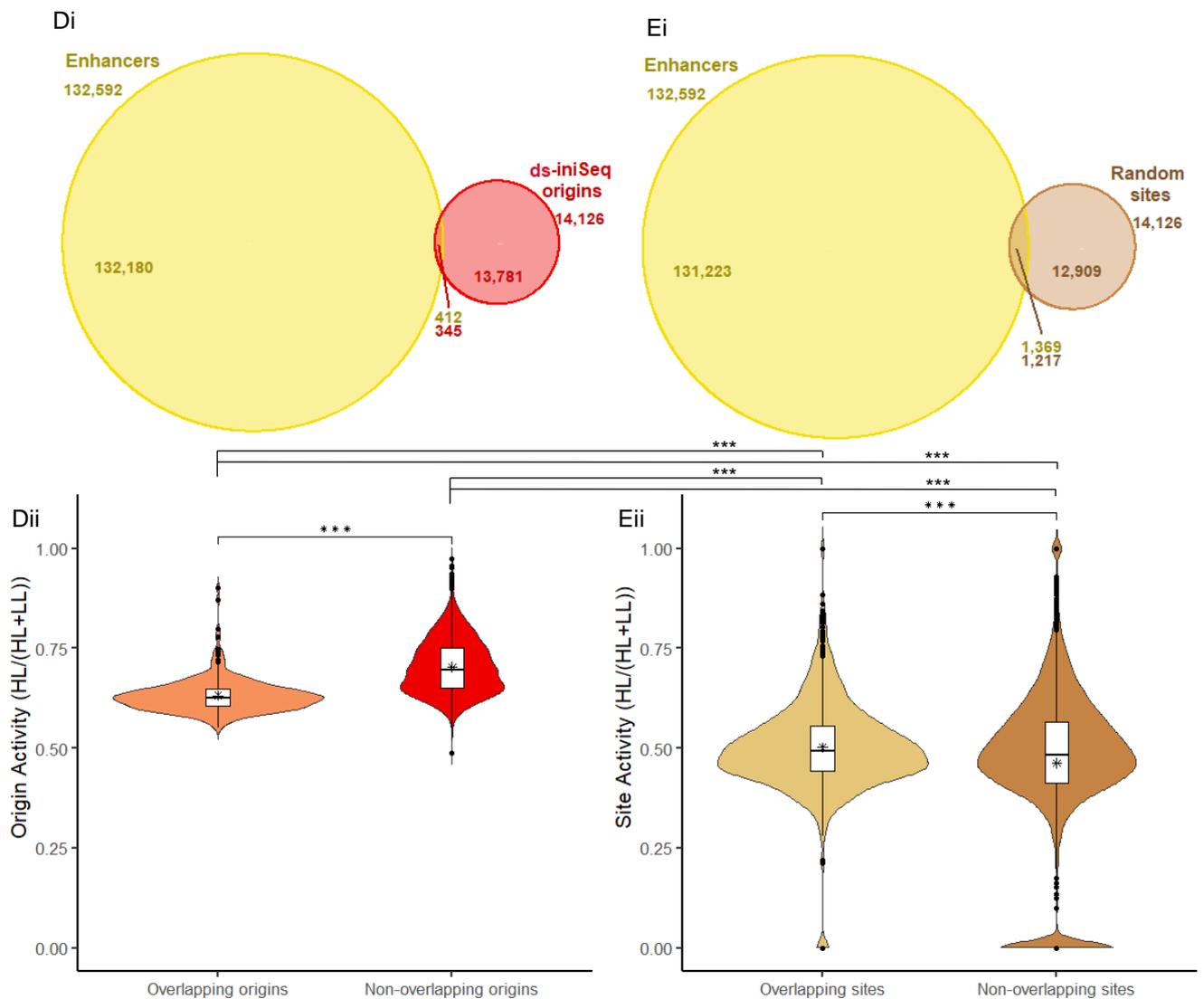


Ci



Cii





**Figure 5.12:** The called origins found in replicate 1 that overlap with the called origins in replicates 2 and 3 (red) were compared to the annotated promoters (apple green) and enhancers (gold) (regulatory features; ensembl). (Ai) The number of ds-iniSeq origins that overlap with promoters (red) and vice versa (apple green). (Aii) The origin activity of the ds-iniSeq origins that do (overlapping origins) and do not (non-overlapping origins) overlap with promoters. The mean is indicated with an \* and a student's T-test was conducted; \*\*\* indicates  $p < 0.001$ . (Bi) The number of ds-iniSeq random sites that overlap with promoters (ochre) and vice versa (apple green). (Bii) The origin activity of the ds-iniSeq random sites that do (overlapping sites) and do not (non-overlapping sites) overlap with promoters. The mean is indicated with an \* and a student's T-test was conducted; \*\*\* indicates  $p < 0.001$ . (Ci) The 3 way overlap of CpG islands (grey) and promoters (apple green) with the ds-iniSeq origins (red). (Cii) The origin activity of the ds-iniSeq origins that; did not overlap with CpG islands or promoters ("non-overlapping"); overlapped with only CpG islands; overlapped with only promoters; and that overlapped with both CpG islands and promoters. The means are indicated with an \*. An ANOVA and subsequent Tukey's post-hoc test were performed to assess significance; the Tukey's test results are shown on the plot and \*\*\* indicates  $p < 0.001$  and \* indicates  $p < 0.05$ . (Di) The number of ds-iniSeq origins that overlap with enhancers (red) and vice versa (gold). (Dii) The origin activity of the ds-iniSeq origins that do (overlapping origins) and do not (non-overlapping origins) overlap with enhancers. The mean is indicated with an \* and a student's T-test was conducted; \*\*\* indicates  $p < 0.001$ . (Ei) The number of ds-iniSeq random sites that overlap with enhancers (ochre) and vice versa (gold). (Eii) The origin activity of the ds-iniSeq random sites that do (overlapping sites) and do not (non-overlapping sites) overlap with enhancers. The mean is indicated with an \* and a student's T-test was conducted; \*\*\* indicates  $p < 0.001$ . An ANOVA and subsequent Tukey's post-hoc test was performed to assess significance between ds-iniSeq origins and their corresponding random sites; the Tukey's test results are shown on the plot; \*\*\* indicates  $p < 0.001$ .

The overlap analysis of the ds-*iniSeq* origins and promoters (Fig.5.12Ai) showed a high degree of overlap; >74% of ds-*iniSeq* origins overlapped with ~30% of the 35,191 promoters. By contrast, only 5.2% of ds-*iniSeq* RS overlapped with ~2.3% of promoters (Fig.5.12Bi). Again, these data demonstrated the specific enrichment of ds-*iniSeq* origins at promoters, indicating that they did not do so by chance. Once more, this was consistent with prior research which found an association with promoters and origins (23,30).

The relative activities of those ds-*iniSeq* origins that overlapped with promoters were significantly higher than that of non-overlapping origins (Fig.5.12Aii). The distribution of the non-overlapping origin activities was almost exclusively at lower activity values, whereas the overlapping origins activities were more evenly distributed across a range of higher relative origin activities.

As before, the relative activities of those ds-*iniSeq* RS that did and did not overlap promoters (Fig.5.12Bii) were significantly lower than that of the ds-*iniSeq* origins that did and did not overlap promoters.

The presence of a promoter at a ds-*iniSeq* origin, correlated with increased origin activity, highlighting the potential of promoters to increase origin activity and efficiency.

Given the strong colocalisation of promoters with CGIs (26) and the strong association of CGIs with ds-*iniSeq* origins possessing high activities (Fig.5.9), these two elements must be considered together. It was impossible to distinguish the influence of promoters from CGIs on origin activity from the data Fig.5.12A. I therefore analysed ds-*iniSeq* origins that associated with promoters and/or CGIs (Fig.5.12C).

Most of the ds-*iniSeq* origins overlapped with both CGIs and promoters (68.1%). The next largest overlap of the origins was with the CGIs only (7.5%), followed by the overlap with the promoters only (6.1%). Approximately 18% of these origins overlapped with neither (Fig.5.12Ci).

The ds-*iniSeq* origins relative activities of those that did not overlap with either feature possessed the significantly lowest origin activities with an accumulation of origins at lower activities (Fig.5.12Cii). Both non-overlapping and promoter only overlapping origins were highly significantly less active than origins overlapping CGIs only ( $p < 2 \times 10^{-16}$ ) and both CGIs and promoters ( $p < 2 \times 10^{-16}$ ). The CGI only overlapping origins possessed the second highest relative activity and a wider distribution of activities when compared to the non-overlapping and promoter only overlapping origins. Those origins that overlapped with both CGIs and promoters had the statistically highest relative origin activities, with a fairly even and wide distribution.

Yet again, this highlighted the association with CGIs is a dominant feature for increased origin activity. There was an additional effect on origin activity when they associated with CGIs and promoters, highlighting a potential role for CGI-promoters in specifying DNA replication origins. The prominence of CGI-promoters as a feature that promotes origin firing has been documented in mouse embryonic stem cells, which showed that the origins colocalised with CGI-promoters were the most efficient (23); together with my data, these indicate evolutionary conservation of the potential role of CGI-promoters in origin firing.

CGI-promoters are well known for their role in transcription where they adopt a transcriptionally permissive state, which is believed to generate a chromatin structure with a more accessible environment to enable RNA polymerase II binding (41). This state is possibly applicable to replication origins and the replication machinery can assemble at the more accessible sites produced by the CGI-promoter induced permissive state.

The overlap analysis of ds-iniSeq origins and enhancers (Fig.5.12Di) showed that only 2.4% of ds-iniSeq origins overlapped with ~0.3% of the 132,592 enhancers. Conversely, 8.6% of ds-iniSeq RS overlapped ~1% of the enhancer (Fig.5.12Ei). Intriguingly, the ds-iniSeq RS were present at more enhancer sites than the corresponding ds-iniSeq origins, which may suggest that enhancers were being actively excluded from these origins. This finding is in disagreement with previous literature which asserts that replication origins colocalise with enhancers (22). However, I find that the almost exclusion of ds-iniSeq origins at enhancers reasonably unsurprising, as enhancers can be great distances from gene bodies and in intergenic DNA (42), which colocalise with fewer origins (1) at lower origin activities (Fig.5.11).

The relative activities of ds-iniSeq origins that overlapped enhancers were significantly lower than that of those that did not overlap enhancers (Fig.5.12Dii). The distribution of origin activities of overlapping origins was almost exclusively at lower activity values, whereas the non-overlapping origins activities were more evenly distributed across the range of higher activities. This was the inverse relationship of that seen by the overlap analysis of ds-iniSeq origins with promoters and the other genomic features examined here, suggesting that enhancers were associated with less efficient replication origins and thus origins with lower probabilities of firing. Enhancers may be consistent with or generate an environment which is less favourable for origin firing/initiation.

Finally, the relative activities of ds-iniSeq RS that did and did not overlap with enhancers (Fig.5.12Eii), showed an almost identical result to the analysis of ds-iniSeq RS with the promoters (Fig.5.12Bii). The relative activities of ds-iniSeq RS that did and did not overlap with enhancers were significantly lower than the activities of the ds-iniSeq origins overlapping and not overlapping enhancers.

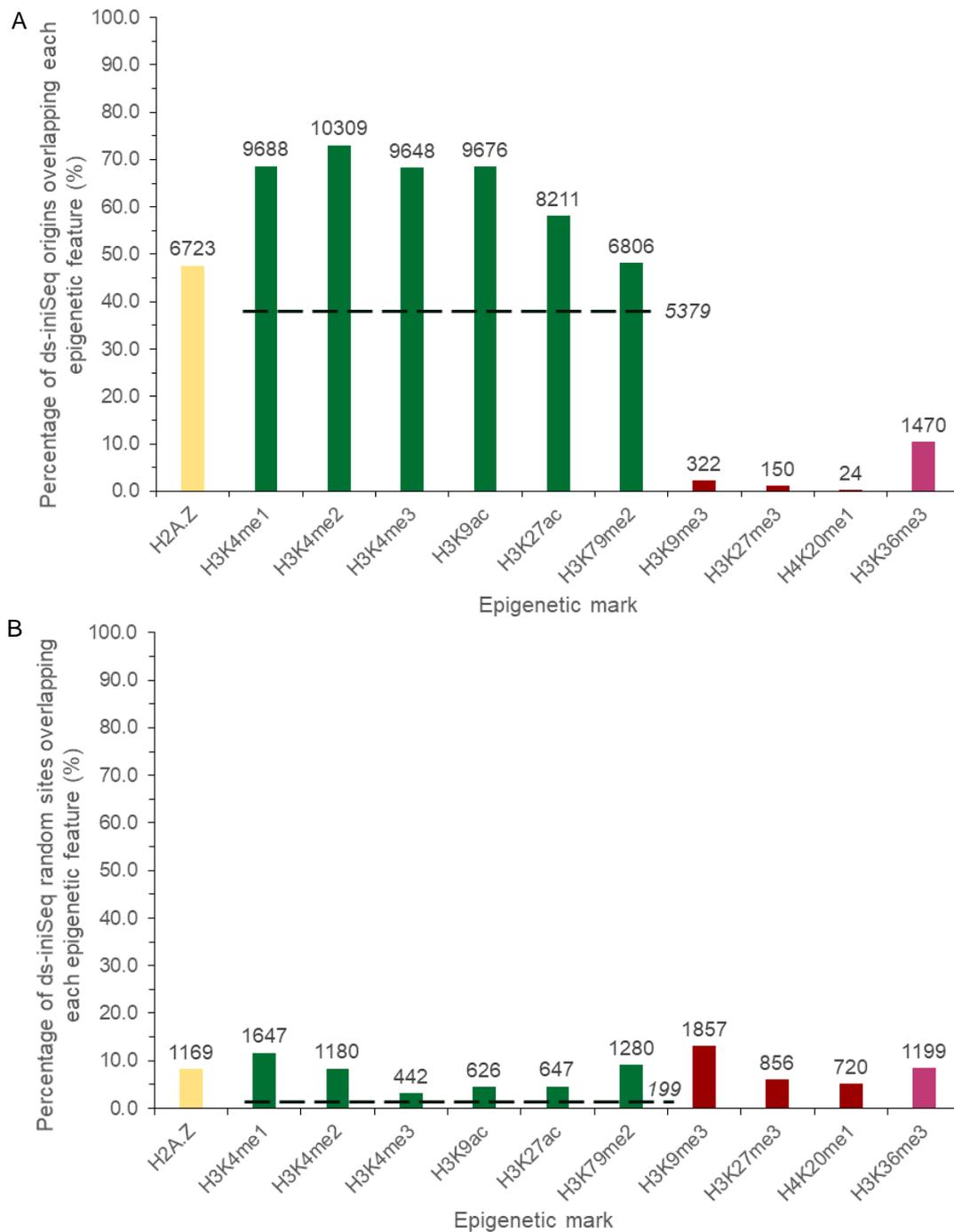
Together, all these data (Fig.5.9-5.12) appeared to show that the degree of overlap between a genomic feature and the ds-iniSeq origins was an indicator of the origin activity. Features that frequently overlapped ds-iniSeq origins were associated with greater overall origin activities/higher probabilities for origin initiation. Furthermore, genomic features excluded at origins were associated with lower activities/probabilities of origin firing.

### *5.2.9 Comparison to epigenetic features*

The literature has reported multiple epigenetic histone marks and isoforms that associate with human DNA replication origins, in both early and late replication. On ENCODE, there was ChIP-seq data for many of these identified epigenetic features conducted on the human colorectal epithelial carcinoma cell line HCT116, which I decided was sufficiently similar to EJ30s to be used for my subsequent analyses.

The ChIP-seq data for the epigenetic features associated with early replication origins were H2A.Z, H3K4me1/2/3, H3K9ac, H3K27ac and H3K79me2. The ChIP-seq data for the histone marks associated with late replication origins were H3K9me3 and H3K27me3. I also included H4K20me1, which has been associated with replication origins. H4K20me1 is believed to act as a chromatin template for subsequent di- and tri-methylation of H4K20. H4K20me3 is associated with late-firing origins and the insurance of correct heterochromatin replication timing (43–47). For the purposes of this analysis, I included H4K20me1 as a Late Firing Origin Associated (LFOA) mark, as a proxy for H4K20me3. The active transcription histone mark H3K36me3 has been suggested as a potential Early Firing Origin Associated (EFOA) mark, but the literature remains ambiguous with respect to this role. I included H3K36me3 in the analysis as an active transcription mark.

I performed overlap analyses to assess whether these epigenetic features colocalise with the ds-iniSeq origins, and if this correlated with different ds-iniSeq origin activities (Fig.5.13). As a caveat, these ChIP-seq data were conducted in a different cell line and may have lineage differences in the distribution of these epigenetic features. The subsequent analysis must be considered with that in mind.



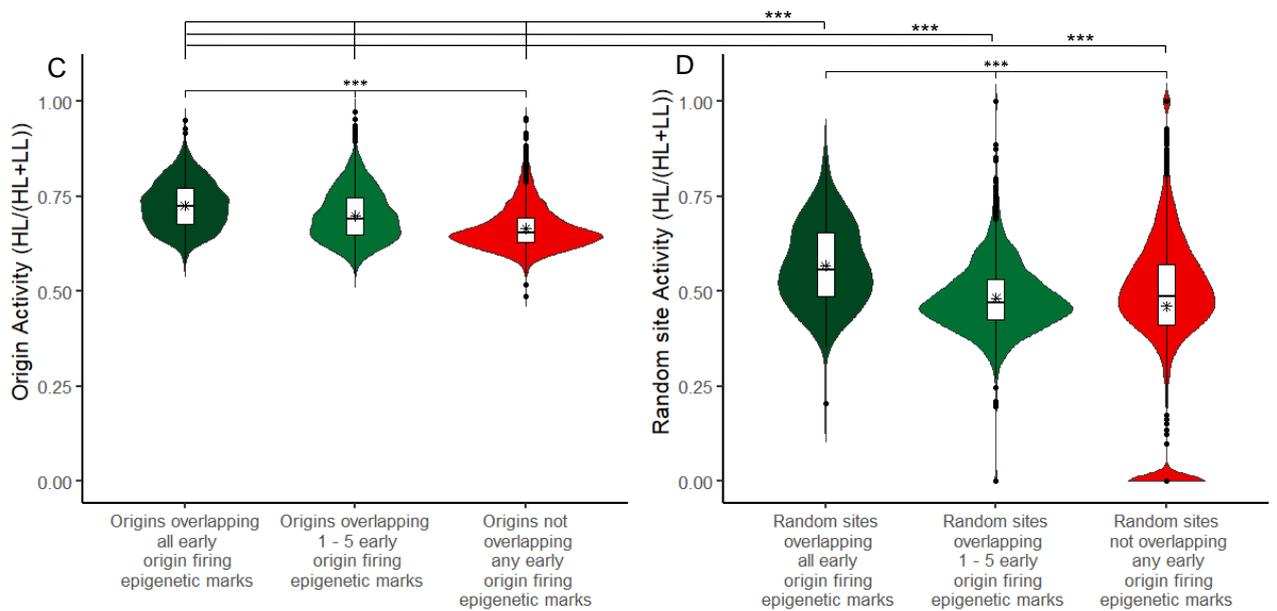
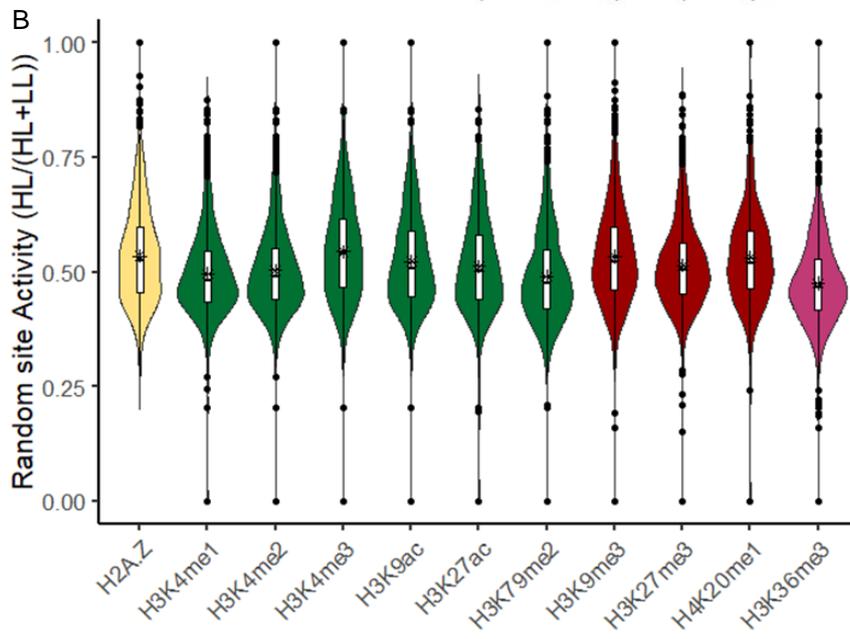
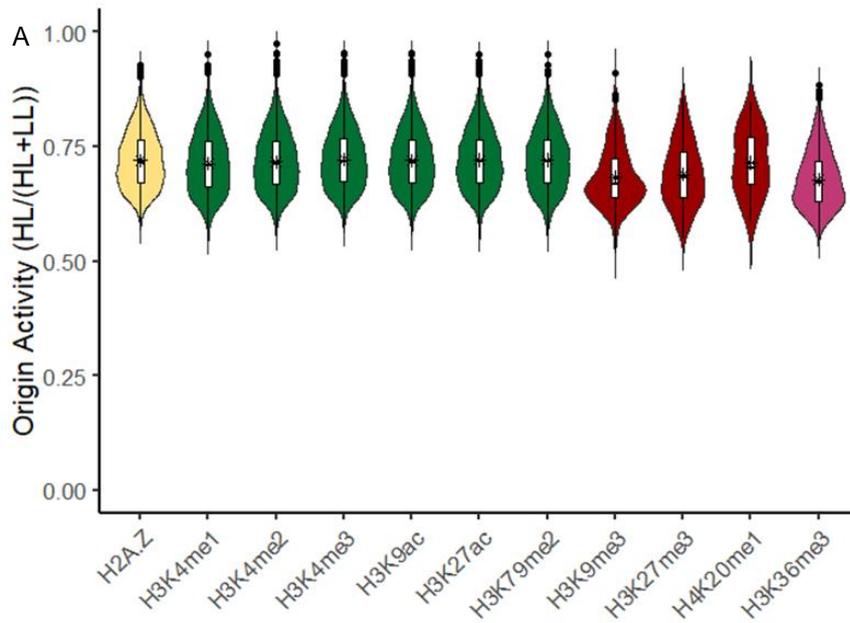
**Figure 5.13:** The called origins found in replicate 1 that overlap with the called origins in replicates 2 and 3 (red) were compared to CHIP-seq data for the HCT116 cell line (ENCODE) for the histone isoform H2A.Z (associated with DNA replication; pale yellow) and histone marks associated with early (dark green) and late (dark red) firing replication origins, and one histone mark associated with transcriptional activity (pink). (A) The percentage of ds-iniSeq origins that were found to overlap with these epigenetic features were shown and the total number of origins overlapping those features were shown above the appropriate column. The percentage and number of origins that were found to overlap all histone marks associated with early firing origins were indicated with a very dark green dashed line. (B) The percentage of ds-iniSeq random sites that were found to overlap with these epigenetic features were shown and the total number of sites overlapping those features were shown above the appropriate column. The percentage and number of sites that were found to overlap all histone marks associated with early firing origins were indicated with a very dark green dashed line.

The percentages of ds-iniSeq origins overlapping the above epigenetic features (and the raw origin numbers) are shown in Fig.5.13A. Nearly 50% of the ds-iniSeq origins overlapped with the H2A.Z histone isoform. The ds-iniSeq origins strongly colocalised with the previously identified EFOA histone marks, H3K4me1 (68.6%), H3K4me2 (73.0%), H3K4me3 (68.3%), H3K9ac (68.5%), H3K27ac (58.1%) and H3K79me2 (48.2%). I also found that 38.1% ds-iniSeq origins overlapped with all 6 EFOA marks (dashed line). Conversely, very few ds-iniSeq origins overlapped with the LFOA histone marks, H3K9me3 (2.3%), H3K27me3 (1.1%) and H4K20me1 (0.2%). Finally, few ds-iniSeq origins overlapped with the active transcription marker, H3K36me3 (10.4%).

For comparison, I performed an overlap analysis of the corresponding ds-iniSeq RS with the same epigenetic features (Fig.5.13B). Only 8.3% of ds-iniSeq RS overlapped with the histone isoform H2A.Z. Few ds-iniSeq RS overlapped with the EFOA marks, H3K4me1 (11.7%), H3K4me2 (8.4%), H3K4me3 (3.1%), H3K9ac (4.4%), H3K27ac (4.6%) and H3K79me2 (9.1%). Only 1.4% of ds-iniSeq RS overlapped with all 6 EFOA histone marks. The ds-iniSeq RS also showed a low level of overlap with the LFOA histone marks, H3K9me3 (13.1%), H3K27me3 (6.1%) and H4K20me1 (5.1%) and the active transcription marker, H3K36me3 (8.5%).

These data (Fig.5.13) showed a specific enrichment of ds-iniSeq origins at EFOA epigenetic features and an enrichment of ds-iniSeq origins at sites occupied by all 6 EFOA histone marks, over and above random chance. Conversely, ds-iniSeq origins were almost excluded at LFOA histone marks but there was no enrichment or depletion of ds-iniSeq origin at transcription-associated H3K36me3 sites.

I investigated the relative activities of the ds-iniSeq origins and RS that overlapped with each of these epigenetic features (Fig.5.14).



**Figure 5.14:** The called origins found in replicate 1 that overlap with the called origins in replicates 2 and 3 (red) were compared to ChIP-seq data for the HCT116 cell line (ENCODE) for the histone isoform H2A.Z (associated with DNA replication; pale yellow) and histone marks associated with early (dark green) and late (dark red) firing replication origins, and one histone mark associated with transcriptional activity (pink). (A) The violin plot of origin activity of the ds-iniSeq origins that overlapped these epigenetic marks and (B) the violin plot of the site activity of the ds-iniSeq random sites that overlapped with these epigenetic features. (C) The origin activity of the ds-iniSeq origins that; did not overlap any early replication associated histone marks, overlapped between 1 and 5 early replication associated histone marks and overlapped all 6 early replication associated histone marks. (D) The site activity of the ds-iniSeq random sites that; did not overlap any early replication associated histone marks; overlapped between 1 and 5 early replication associated histone marks; and overlapped all 6 early replication associated histone marks. (C & D) The means are indicated with an \*. An ANOVA and subsequent Tukey's post-hoc test were performed to assess significance; the Tukey's test results are shown on the plot and \*\*\* indicates  $p < 0.001$ .

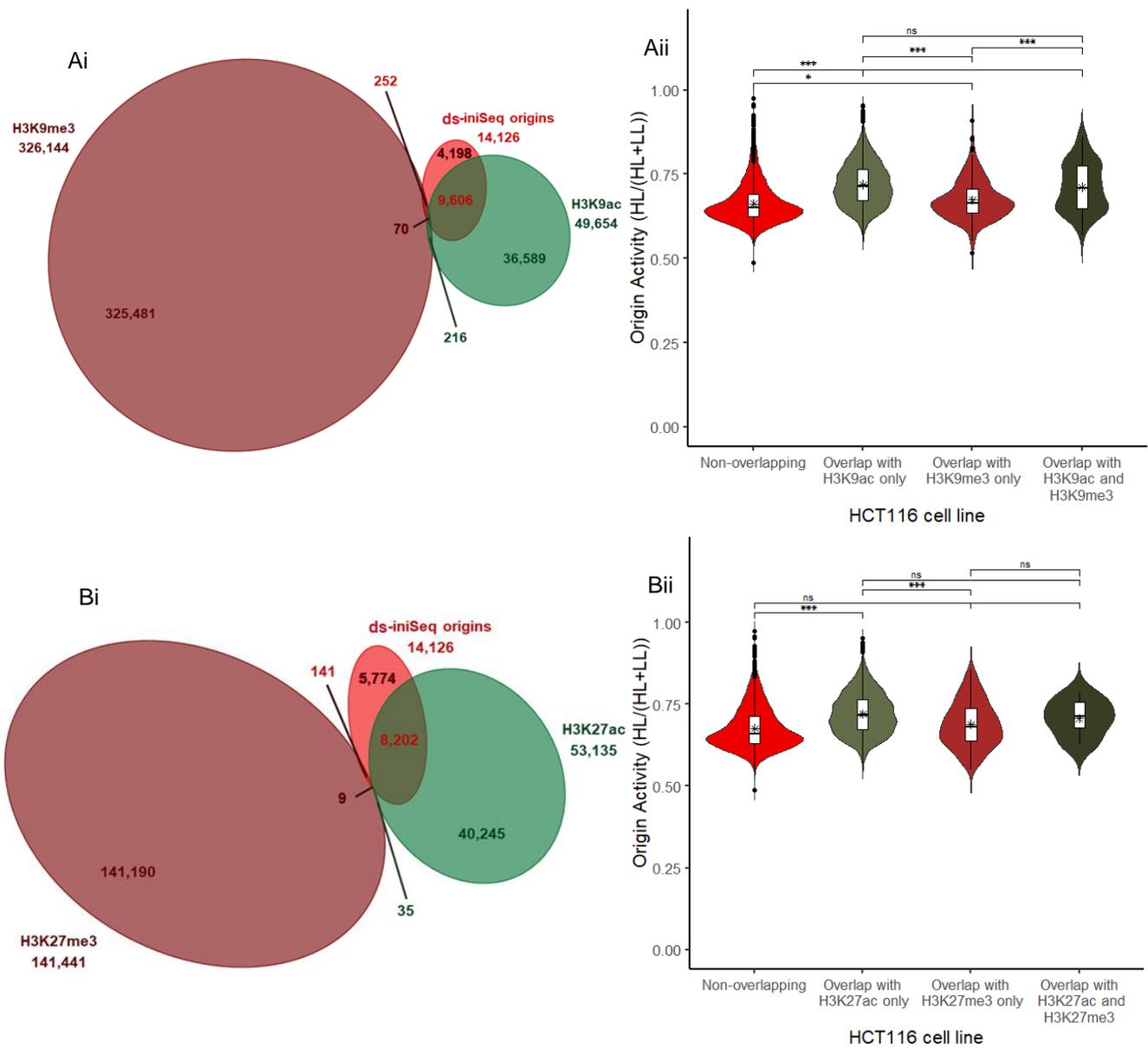
The relative activities of the ds-iniSeq origins that overlapped with EFOA epigenetic features all possessed higher activities and almost identical, evenly spread distributions (Fig.5.14A). The relative activities of the H3K9me3 overlapping ds-iniSeq origins were much lower and had an accumulation at the lowest activities, whereas the relative activities of the H4K20me1 overlapping ds-iniSeq origins were similar to the activities of the EFOA epigenetic feature overlapping ds-iniSeq origins. The ds-iniSeq origins that overlapped with H3K27me3 sites possessed lower origin activities than those overlapping each EFOA epigenetic feature and H4K20me1, but higher than the H3K9me3 overlapping origins activities. The distribution of the H3K27me3 overlapping origin activities were similar to that of the origins overlapping EFOA epigenetic features, but slightly shifted towards lower activities. The relative activities of the ds-iniSeq origins overlapped with H3K36me3 sites were similar to the H3K9me3 overlapping ds-iniSeq origin activities but possessed a less distinct accumulation of lower activities.

The ranges and distributions of relative activities of the ds-iniSeq RS (Fig.5.14B) varied for origins overlapped with each epigenetic feature, but there was no obvious bias in site activities of those overlapping EFOA vs LFOA epigenetic features. This, again, intimates that the associations of ds-iniSeq origins with epigenetic marks, associated with replication, are specific over and above chance.

To determine whether the number of EFOA marks that colocalise with an origin is associated with origin activity, I assessed the relative activities of the ds-iniSeq origins that overlapped with all 6 EFOA histone marks, 1-5 EFOA marks or no EFOA marks (Fig.5.14C). Those origins that overlapped with all 6 EFOA marks possessed the significantly highest origin activities. Those origins that overlapped with 1-5 EFOA marks possessed significantly higher activities than the origins overlapping no EFOA marks but significantly lower origin activities than the origins overlapping all 6 EFOA marks. Those origins overlapping no EFOA marks possessed the significantly lowest activities, with an accumulation of lower origin activities.

The relative activities of ds-*iniSeq* RS overlapping all 6 EFOA histone marks, 1-5 EFOA histone marks or no EFOA marks (Fig.5.14D) were all significantly lower than that of the ds-*iniSeq* origins that overlap all 6 EFOA, 1-5 EFOA or no EFOA histone marks. Therefore, the increased activity in of the ds-*iniSeq* origins was over and above chance.

The H3K9 and H3K27 histone positions have been found in both the acetylated and trimethylated states in EFOA and LFOA histone marks, respectively. Therefore, I further investigated the relationship between these marks (HCT116 cell line) and the ds-*iniSeq* origins they overlapped with (Fig.5.15).



**Figure 5.15:** The H3K9 and H3K27 histone positions are both acetylated and trimethylated which are associated with early and late firing replication origins respectively. (Ai) The 3 way overlap of the early replication associated H3K9ac (HCT116 cell line; dark green) and the late replication associated H3K9me3 (HCT116 cell line; dark red) with ds-iniSeq origins (red). (Aii) The origin activity of the ds-iniSeq origins that; did not overlap with H3K9ac or H3K9me3 (“non-overlapping”), overlapped with only H3K9ac, overlapped with only H3K9me3 and that overlapped with both H3K9ac and H3K9me3. (Bi) The 3 way overlap of the early replication associated H3K27ac (HCT116 cell line; dark green) and the late replication associated H3K27me3 (HCT116 cell line; dark red) with ds-iniSeq origins (red). (Bii) The origin activity of the ds-iniSeq origins that; did not overlap with H3K9ac or H3K27me3 (“non-overlapping”), overlapped with only H3K9ac; overlapped with only H3K27me3; and that overlapped with both H3K27ac and H3K27me3. (All ii) The means are indicated with an \*. An ANOVA and subsequent Tukey’s post-hoc test were performed to assess significance; the Tukey’s test results are shown on the plot and \*\*\* indicates  $p < 0.001$ , \* indicates  $p < 0.05$  and ns indicates “not significant”.

The three-way overlap analysis of the ds-iniSeq origins with H3K9ac and/or H3K9me3 sites showed (Fig.5.15Ai) that very few H3K9ac sites overlapped with H3K9me3 sites (216). The ds-iniSeq origins predominantly overlapped with only the H3K9ac sites (~68%) and very few overlapped with H3K9me3 (1.8%), with even fewer ds-iniSeq origins overlapped with both (~0.5%).

The ds-iniSeq origins that did not overlap with either H3K9ac or H3K9me3 had the significantly lowest relative activities with an accumulation of lower activities (Fig.5.15Aii). A similar distribution, with an accumulation of lower activities, was seen in those origins that overlapped H3K9me3 only, but these ds-iniSeq origin activities were significantly higher ( $p=0.24$ ).

The ds-iniSeq origins that overlapped with H3K9ac only possessed the highest activities which were significantly greater than non-overlapping and H3K9me3 only overlapping origins (Fig.5.15Aii). The origins overlapping both H3K9ac and H3K9me3 were not significantly less active than the H3K9ac only overlapping origins but were significantly more active than the ds-iniSeq origins in the other overlap conditions.

As with H3K9 and in line with expectation, negligible numbers of H3K27ac sites overlapped with H3K27me3 sites (35) (Fig.5.15Bi). The ds-iniSeq origins predominantly overlapped with only H3K9ac sites (58%), ~1% overlapped with H3K9me3 and only 0.06% overlapped with both H3K9ac and H3K9me3.

The relative activities of the ds-iniSeq origins that did not overlap with H3K27ac or H3K9me3 (Fig.5.15Bii) were significantly lower than origins overlapping H3K27ac only, but not significantly different from the ds-iniSeq origins in the other overlap conditions. The ds-iniSeq origins overlapping only H3K27ac had the highest relative activities and a distribution that was almost identical to the H3K9ac overlapping origins (Fig.5.15Aii). The H3K27ac overlapping origins' activities were significantly higher than the H3K27me3 only overlapping origins, but not significantly different from the H3K27ac and H3K27me3 overlapping origins.

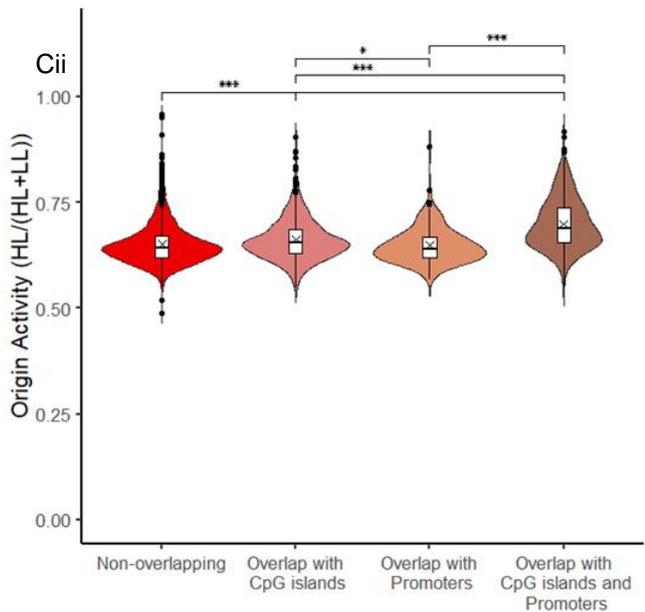
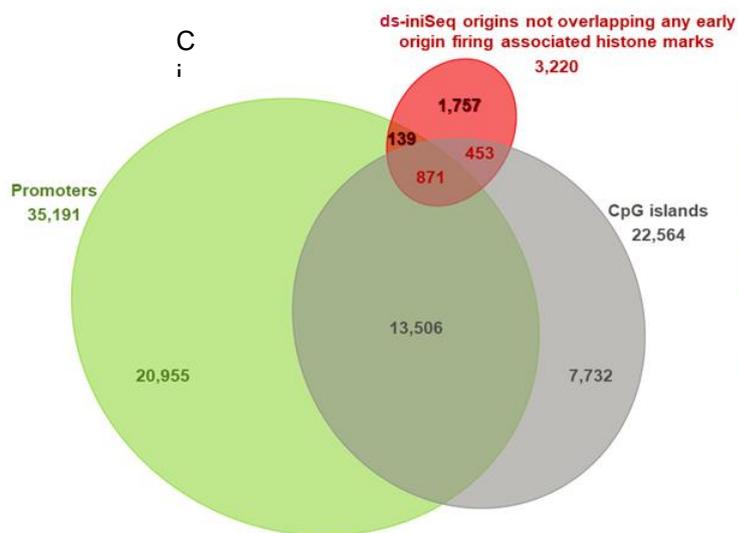
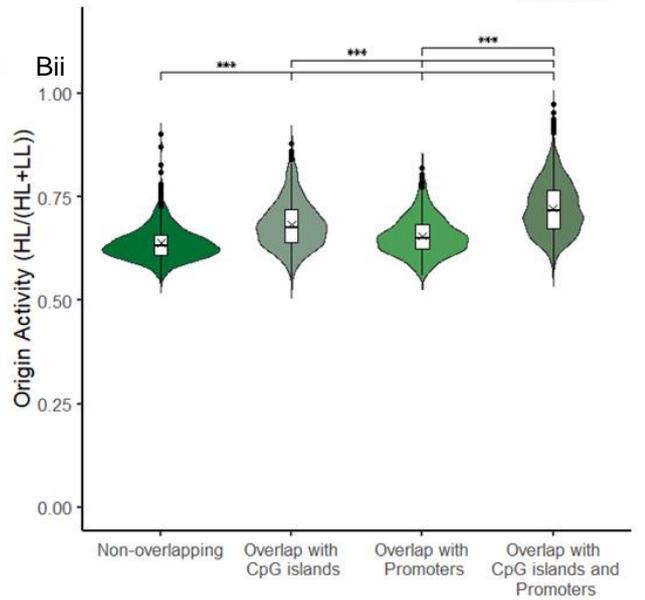
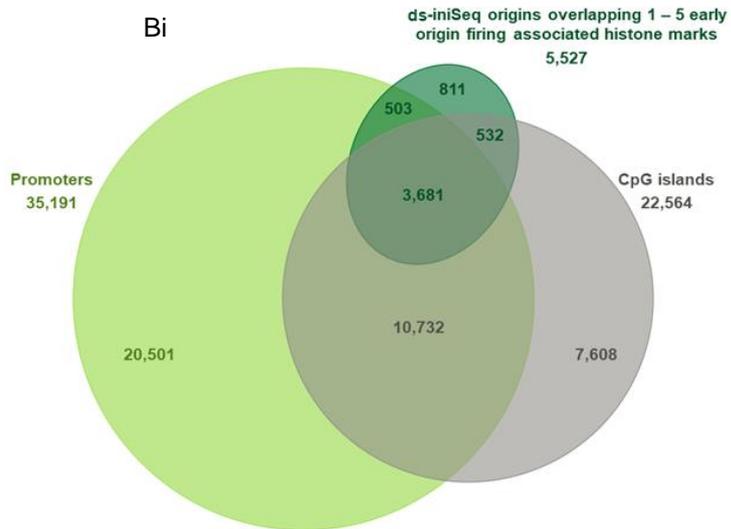
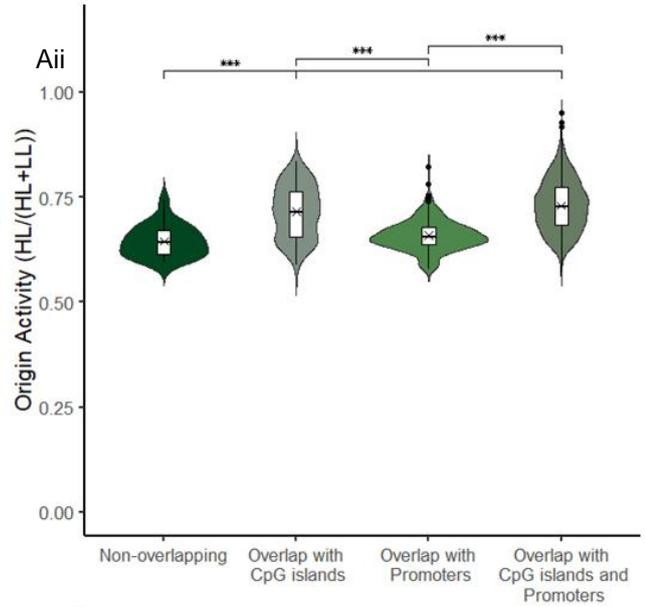
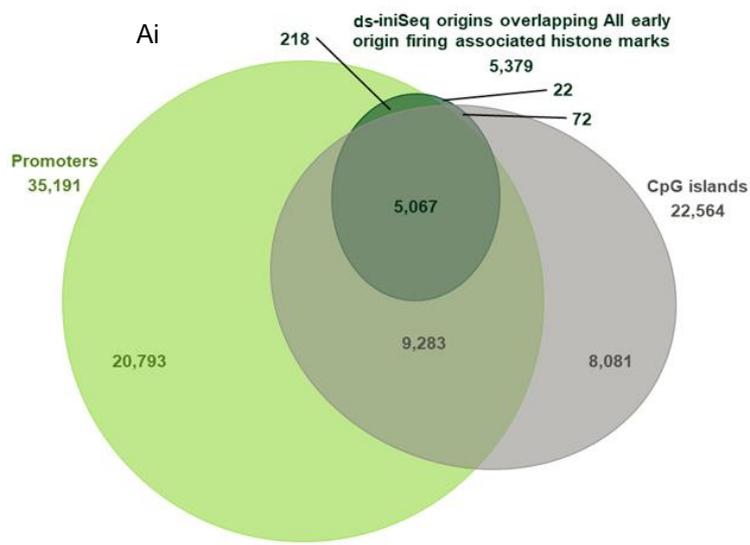
The ds-iniSeq origins that overlapped with the H3K27me3 sites possessed lower average activity. The 9 ds-iniSeq origins that overlapped with both H3K27ac and H3K27me3 sites possessed high origin activities but were not significantly different to any of the other origins.

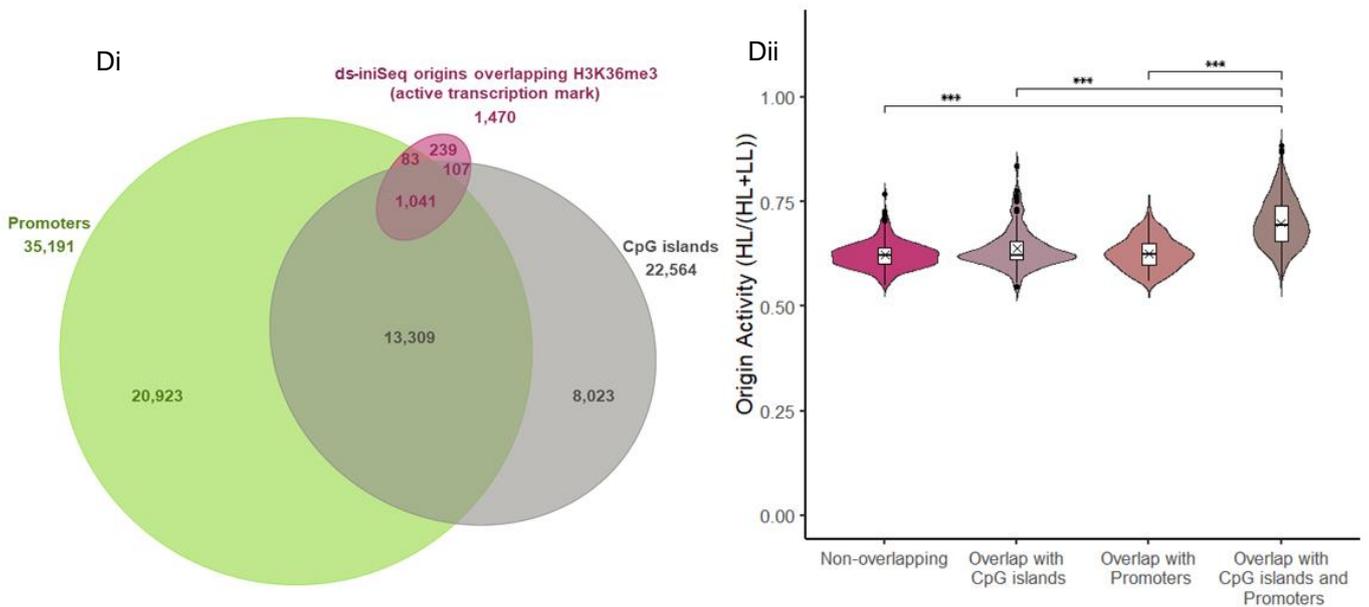
These data show that ds-iniSeq origins were enriched at H3K9ac (Fig.5.15Ai) and H3K27ac sites (Fig.5.15Bi) but depleted at their methylated counterparts and very few ds-iniSeq origins were found at both acetylated and methylated sites. It also shows that acetylated versions of H3K9 (Fig.5.15Aii) and H3K27 (Fig.5.15Bii) colocalised with more active ds-iniSeq origins and when ds-iniSeq origins overlapped both acetylated and methylated sites (for both H3K9 and H3K27), the effect, on origin activity, of the acetylation was dominant.

The above analyses were also carried out on H3K9ac/me3 and H3K27ac/me3 ChIP-seq data conducted in the human prostate epithelial adenocarcinoma cell line, PC3 (appendix Fig.A5.3). The results were similar to those found in the HCT116 cell line.

#### *5.2.10 Interplay between genomic and epigenetic features*

Finally, there may be a relationship/interaction between genomic and epigenetic features in origin specification and activation. I found that ds-iniSeq origins strongly correlated with CGIs and CGI-promoters and these origins possessed the highest activities. Similarly, highly active ds-iniSeq origins strongly colocalised with EFOA histone marks. I investigated the potential impact of CGIs, promoters and epigenetic features on ds-iniSeq origin activity. I performed an overlap and relative activity analysis of ds-iniSeq origins that were found at, all 6 EFOA marks, some (1-5) EFOA marks, no EFOA marks, or the active transcription mark H3K36me3, with CGIs, promoters, and both (Fig.5.16).





**Figure 5.16:** The relationship between early firing origin associated histone marks, CpG islands and promoters, and their impact on ds-iniSeq origin. (Ai) The 3 way overlap of the the ds-iniSeq origins overlapping with all 6 early origin firing replication associated histone marks (HCT116 cell line; very dark green) and CpG islands (grey) and promoters (apple green). (Aii) The origin activity of the ds-iniSeq origins that; did not overlap with CpG islands or promoters (“non-overlapping”); overlapped with only CpG islands; overlapped with only promoters; and that overlapped with both CpG islands and promoters. (Bi) The 3 way overlap of the the ds-iniSeq origins overlapping with some (1-5) early origin firing replication associated histone marks (HCT116 cell line; dark green) and CpG islands (grey) and promoters (apple green). (Bii) The origin activity of the ds-iniSeq origins that; did not overlap with CpG islands or promoters (“non-overlapping”); overlapped with only CpG islands; overlapped with only promoters; and that overlapped with both CpG islands and promoters. (Ci) The 3-way overlap of the ds-iniSeq origins overlapping with no early origin firing replication associated histone marks (HCT116 cell line; dark red) and CpG islands (grey) and promoters (apple green). (Cii) The origin activity of the ds-iniSeq origins that; did not overlap with CpG islands or promoters (“non-overlapping”); overlapped with only CpG islands; overlapped with only promoters; and that overlapped with both CpG islands and promoters. (Di) The 3-way overlap of the the ds-iniSeq origins overlapping with active transcription histone mark, H3K36me3 (HCT116 cell line; pink) and CpG islands (grey) and promoters (apple green). (Dii) The origin activity of the ds-iniSeq origins that; did not overlap with CpG islands or promoters (“non-overlapping”); overlapped with only CpG islands; overlapped with only promoters; and that overlapped with both CpG islands and promoters. (All ii) The means are indicated with an “x”. An ANOVA and subsequent Tukey’s post-hoc test were performed to assess significance; the Tukey’s test results are shown on the plot and \*\*\* indicates  $p < 0.001$  and \* indicates  $p < 0.05$ ; all other combinations were not significant.

Of the 5,379 ds-iniSeq origins found at all 6 EFOA marks (H3K4me1/2/3, H3K9ac, H3K27ac and H3K79me2), only 0.41% did not colocalise with CGIs and/or promoters (Fig.5.16Ai). Most of these ds-iniSeq origins overlapped with both CGIs and promoters (94%), and a few additional origins overlapped with either CGIs (1.3%) or promoters (4.1%).

The relative activities of the ds-iniSeq origins, colocalised with all EFOA marks, that did not overlap with CGIs and/or promoters possessed the lowest origin activity, with a moderate accumulation of lower activity origins (Fig.5.16Aii). Those origins that overlapped with promoters only were not significantly more/less active than the non-overlapping origins and a

greater accumulation of lower activity origins. The relative activities of both the non-overlapping and promoter only overlapping origins were highly significantly lower than the activities of those origins that overlap with CGIs and both CGIs and promoters and had no accumulation of lower activity origins. Ds-iniSeq origins overlapping both CGIs and promoters possessed the significantly highest activities when compared to all other origins (Fig.5.16Aii).

Of the 5,527 ds-iniSeq origins found at some (least 1, but not all 6; ie 1-5) EFOA marks, 14.7% did not overlap with CGIs and/or promoters, 67% overlapped with CGIs and promoters, 9.6% overlapped with CGIs only and 9.1% overlapped with promoters only (Fig.5.16Bi). The proportion of the origins associated with 1-5 EFOA marks that overlapped with both genomic features decreased compared to the origins found at all EFOA histone marks.

The relative activities of ds-iniSeq origins associated with 1-5 EFOA marks that did not overlap with either feature possessed the significantly lowest origin activity, with an accumulation of lower activity origins (Fig.5.16Bii). The ds-iniSeq origins that overlapped with promoters only possessed the second lowest activities, which was significantly greater than the non-overlapping origins. The relative activities of the ds-iniSeq origins that overlapped CGIs only were significantly higher than the non-overlapping and promoter only overlapping origins and had no accumulation at the lower activities. The relative activities of origins overlapped with both CGIs and promoters were statistically the highest and had no accumulation of origins at lower activities. The relative activities of the origins associated with 1-5 EFOA marks, that overlapped with both genomic features decreased compared to the origins found at all EFOA histone marks.

Of the 3,220 ds-iniSeq origins found at no EFOA marks, the largest proportion of 55% did not overlap with CGIs and/or promoters (Fig.5.16Ci). The fewest of these origins overlapped with promoters only (4.3%), followed by those origins overlapped with CGIs only (14.1%). Finally, 27.1% of these origins overlapped with both CGIs and promoters. The proportion of the origins associated with no EFOA marks, that overlapped with both genomic features further decreased compared to the origins found at all and 1-5 EFOA histone marks.

The relative activities of ds-iniSeq origins, associated with no EFOA histone marks, that overlapped with CGIs and promoters were significantly highest (Fig.5.16Cii) and had the largest and highest distribution, with a slight shift towards lower activities. The non-overlapping origins possessed the statistically lowest activities, closely followed by the promoter only overlapping origins and finally followed by CGIs only overlapping origin; all origins had an accumulation of lower relative origin activities. The activities of the origins

associated with no EFOA marks that overlapped with both genomic features decreased further, compared to the origins found at all and 1-5 EFOA histone marks.

Taken together, Fig.5.16A-C demonstrated a relationship between EFOA marks with CGIs, promoters and CGI-promoters that impacted replication origin firing. The greater the number of EFOA marks colocalised with ds-iniSeq origins, the greater the degree of overlap CGIs and/or promoters. Notably, the association of ds-iniSeq origins with CGI-promoters were substantially reduced as fewer of EFOA marks colocalised with the ds-iniSeq origins.

In origins overlapping some or all the EFOA marks, an association with CGIs only showed increased relative origin activities when compared to that of non-overlapping and promoter-only overlapping origins. However, the impact of CGIs on relative origin activity was dramatically reduced when no EFOA marks were present. When origins colocalised with CGI-promoters, they were more active than any other grouping, irrespective of the number of EFOA marks present. Nevertheless, the relative activity of origins at CGI-promoters decreased as the number of EFOA marks at origins reduced. These findings appeared to indicate a relationship between the epigenetic and genomic features, by illustrating the influence of CGIs and CGI-promoters, in conjunction with EFOA histone marks, on origin specification and/or activation. It is possible that CGIs/CGI-promoters may be involved in directing the epigenetic environment at origins.

Finally, the overlap analysis of the 1,470 ds-iniSeq origins colocalised with H3K36me3, with CGI and promoters showed that most origins overlapped with both CGIs and promoters (70.8%) (Fig.5.16Di). The lowest number of these origins overlapped with promoters only (5.6%), followed by those origins overlapping CGIs only (7.3%). Finally, 16.3% of these origins overlapped with neither genomic feature.

The relative activities of the ds-iniSeq origins, colocalised with H3K36me3, that overlapped with both CGIs and promoters were significantly greatest (Fig.5.16Dii) and were equivalent to the relative activities of the origins associated with no EFOA marks that overlap CGIs and promoters (Fig.5.16Cii). The relative activities of these origins that overlapped neither genomic feature, CGIs only and promoters only were not significantly different but were all highly significantly lower than the activities of the CGI-promoter overlapping origins.

The findings in Fig.5.16A-C indicated a relationship between the CGIs and CGI-promoters and EFOA histone marks that affected origin firing. No such relationship was established with the transcriptional activity mark, H3K36me3. There was a moderate association with CGIs and/or promoters, and CGI-promoter associated origins possessed a higher relative origin activity but there was no increase in relative activity of origins associated with CGIs only when compared to those that overlapped with promoters or neither (Fig.5.16D).

### 5.3 Discussion

In line with previous literature, these data showed that epigenetic marks that were associated with early firing origins colocalised with the ds-*iniSeq* origins (Fig.5.13). A study (conducted on HCT116) identified the binding sites of the Treslin (replication protein) binding partner, MDM2 binding protein (MTBP) and found a large colocalisation with H3K4me1/2/3, H3K27ac, H3K9ac and H2A.Z sites (48,49). This agreed with my data.

The ds-*iniSeq* origins that colocalised with EFOA epigenetic features possessed higher relative origin activities. As these relative ds-*iniSeq* origin activities represent the probability that an origin will fire, these epigenetic features were correlated with higher probabilities of origin firing. The relative activities of the ds-*iniSeq* origins that overlapped with each EFOA histone mark individually were highly similar (Fig.5.14A/B). This may result from the substantial number of origin sites that colocalised with more than one EFOA histone mark (Fig.5.13). I demonstrated that the number of EFOA histone marks found at the ds-*iniSeq* origins had the greatest impact on relative origin activities, where the more histone marks colocalised with ds-*iniSeq* origins, the higher the relative activities/probability of firing (Fig.5.14C/D). I hypothesise that multiple epigenetic marks work together to create an environment that is more conducive with origin firing/activation and as such increases their probability of firing.

By contrast, ds-*iniSeq* origins colocalised with LFOA histone marks showed a much more varied picture. As is expected, there was a very low level of overlap between the ds-*iniSeq* origins and the LFOA marks where they are almost excluded from my ds-*iniSeq* origins (Fig.5.13). Unlike the relationship between degree of overlap and relative origin activity observed with genomic features (ie low colocalisation correlated with low activity), the poor colocalisation of ds-*iniSeq* origins with LFOA marks did not necessarily correlate with low origin activity (Fig.5.14A/B). The variation in relative activities of the few ds-*iniSeq* origins colocalised with LFOA marks may illustrate the differences in the functions/roles of each histone mark in DNA replication.

The histone variant H2A.Z and histone mark H3K79me2 associate with DNA replication origins (28) and my data showed they colocalised with ~50% of ds-*iniSeq* origins (Fig.5.13). The proposed roles for H2A.Z and H3K79me2 tend to imply a more indirect facilitation of DNA replication and origin firing (50–52), which may explain the slightly lower association levels with ds-*iniSeq* origins when compared to other EFOA marks (Fig.5.13).

H3K79me2 is believed to prevent re-replication from occurring and may play a role in preserving replication timing. It's establishment at origins appears to be independent of other epigenetic marks and the prevention of H3K79 dimethylation (by inhibition of DOT1L – only known H3K79 methyltransferase) does not inhibit most cell's ability to replicate and

proliferate (51,52). The ability of cells to replicate in the absence of H3K79me2 may be reflected in its lower association with DNA replication origins identified by ds-iniSeq, as H3K79me2 may not be as crucial to origin firing.

H2A.Z is believed to recruit methyltransferase SUV4-20H1, which in turn dimethylates H4K20 (another histone mark thought to be involved in DNA replication). H4K20me2 then recruits ORC to the initiation sites (50). H2A.Z has also been identified as a potentially important regulator of enhancer activity, which is unlikely to affect my data as I have found no colocalisation of ds-iniSeq origins with enhancers (Fig.5.12D/E). H2A.Z is believed to universally characterise many replication origins as part of a trio of features (H3K4me1 and DNase 1 hypersensitivity) (7). My data showed that other epigenetic features colocalised more substantially with my ds-iniSeq origins, which makes the above assertion less likely (Fig.5.13).

H3K4me1 frequently associates with poised and active promoters (53). H3K4me1 possessed one of the highest degrees of overlap with ds-iniSeq origins, which is consistent with previous literature (Fig.5.13). The assertion that H3K4me1 strongly colocalises with origins as a result of its association with distal enhancers (54) is not supported here. I have demonstrated that the enhancers did not colocalise with ds-iniSeq origins (Fig.5.12D/E). Conversely, I showed that ds-iniSeq origins associated strongly with TSS and promoters (Fig.5.9A/B&12A/B). H3K4me1 are also present at these genomic features (53), which is more consistent with my data. Additionally, H3K4me1, in conjunction with two other histone marks examined here, H3K4me3 and H3K27me3, influence gene transcription (53). DNA replication and transcription should require co-ordination and these three active transcription marks have been found to frequently associate with replication origins (23,28,38,40). Consequently, these three marks may present an avenue for the successful coordination of both processes.

H3K4me3 was highly associated with ds-iniSeq origins, which supports previous literature that demonstrated that H3K4me3 is strongly associated with early firing origins in euchromatin. As with many early-firing associated histone marks, there is a link between H3K4me3 and transcriptional activity. The presence of the active transcription mark, H3K4me3 and the repressive transcription mark, H3K27me3 at a promoter defines a poised promoter, which possesses a bivalent state that can regulate transcription rates depending on the H3K4me3 and H3K27me3 levels at a given promoter (53,55). Along with H3K36me3, H3K4me3 strongly inhibits PRC2 which is the methyltransferase responsible for trimethylating H3K27, whereas H3K27me3 stimulates PRC2 (56–60). As H3K27me3 is associated with late origin firing and H3K4me3 is associated with early origin firing, their partnership at origins may be involved in the regulation of the timing of origin firing by altering the probability of origin firing, in a similar fashion to their proposed control of transcription.

There is, however, limited evidence of this bivalent relationship with ds-iniSeq origins as H3K27me3 is almost excluded, whereas H3K4me3 is highly enriched (Fig.5.13). It may be possible that the impact of the H3K27me3/H3K4me3 bivalent state on origins becomes more apparent later in a replication reaction and these early firing ds-iniSeq origins represent those with only high H3K4me3 levels.

Origins within the same cell line can be classed as either “consistent” or “flexible”. “Consistent” origins fire every cell cycle and are often colocalised with H3K4me3 and H3K9ac. “Flexible” origins can become active or remain dormant during replication, which constitute 80-90% of origins (6,28). In my data, “consistent” origins would be present in all ds-iniSeq replicates and possess higher firing probabilities and thus relative origin activity. Whereas the “flexible” origins would be present in a subset of replicates or in all replicates but possess lower relative origin activities. Ds-iniSeq identifies and determines the relative activity of origins that fire in nuclei from a population of cells. Therefore, a lower relative origin activity would represent fewer instances of a given origin firing in the population, which would explain why, in these data, “flexible” origins could be present in all 3 replicates but may possess lower relative origin activities.

There are also origins that consistently fire in cell lines from many different lineages (shared origins) and origins that are cell-type specific (specific origins) (28). While the influence of histone marks on “consistent” and “flexible” origins require further investigation, the impact of H3K4me3 has been documented in shared and specific origins. H4K3me3 and H3K9ac preferentially colocalise with shared origins but H3K9me3 are weakly associated with these origins. Conversely, H3K9me3 colocalised with origins that consistently fired in a single cell line lineage. The shared origins fired throughout replication, with a bias towards early replication but specific origins predominantly fired in late S-phase (28,34). Therefore, ds-iniSeq origins with lower relative activities may represent cell line specific origins that fire later in S-phase, whereas the higher relative activity ds-iniSeq origins associated with H3K4me3 and H3K9ac may represent shared origins. If this were the case, I would expect a greater colocalisation of origins with H3K9me3 later in a replication reaction. I have developed a method based on ds-iniSeq, where the replication reaction was extended to 3 hours. This density-substitution elongation-site sequencing (ds-eloSeq) (chapter 7) should ultimately be able to identify later firing origins and subsequent elongation sites, later in DNA replication. It could be used to address this hypothesis.

H3K4me2 is increasingly implicated in DNA replication origin specification and activation. It is highly enriched at origins and is necessary and sufficient for normal origin function in *S. cerevisiae*. H3K4me2 is proposed to act with other histone marks to promote stable genomic DNA replication (61). The importance of this histone mark in humans has been highlighted here, where H3K4me2 has the highest enrichment at ds-iniSeq origins. One would have to

conduct a functional study, whereby one inhibits/reverses the di-methylation of H3K4 during a DNA replication reaction in order to assess the impact of this mark on origin and thus establish whether a causal relationship, similar to yeast, exists in humans. Further evidence that corroborates the importance of H3K4me2 at origins comes from both OK-Seq and MTBP-binding site analyses, whereby H3K4me2 was substantially enriched at these sites. Although MTBP-binding sites overlapped with H3K4me1/2/3, H3K27ac, H3K9ac and H2A.Z, they demonstrated the highest colocalisation of 99% with H3K4me2 (7,48). H3K4me2 has been linked to transcriptional activity (62–64) and associated with other EFOA histone marks, H3K27ac, H3K9ac and H3K4me1/3 binding sites. The ds-iniSeq origins demonstrate a high colocalisation with H3K4me2 and other EFOA marks (Fig.5.13). H3K4me2, along with H3K27ac, is enriched at ORC binding sites and are highly accurate predictors of ORC2 binding (4,65). This highlights the complex interaction with these other histone marks, DNA replication and transcription.

H3K9ac and H3K9me3 are mutually exclusive marks, where the acetylation is associated with early firing origins, and the tri-methylation is associated with late firing origins (28,66,67). My data supports these observations, as ds-iniSeq origins represented earlier replicating sites and were highly enriched at H3K9ac sites but almost excluded at H3K9me3; H3K9ac and H3K9me3 showed negligible colocalisation (Fig.5.15A). Ds-iniSeq origins colocalised with H3K9ac were highly active, indicating a greater probability of firing, which is more consistent with firing earlier. Conversely, those ds-iniSeq origins that overlapped with H3K9me3 possessed much lower relative origin activities, indicating a lower probability of firing, which would be more consistent with origin firing later in S-phase. Furthermore, the contrasting relative activities of the ds-iniSeq origins associated with H3K9ac/me3 supports mutually exclusive/antagonistic relationship between the two marks during DNA replication and origin firing.

H3K9ac and H3K9me3 have been proposed as antagonistic co-ordinators of DNA replication and transcription. H3K9ac is associated with transcriptional activity; it has been found to associate with active TSS and at ascending OK-Seq segment borders. Whereas H3K9me3 is associated with transcriptional silencing and heterochromatin as a compaction marker. The Dfb4 promoter exemplifies this co-regulation as it showed that transcription is downregulated when origin firing is high; both processes were inversely regulated. Additionally, H3K9ac was most prevalent at the Dfb4 promoter when transcription was highest, whereas H3K9me3 was most prevalent during and just after DNA replication (7,66–68).

My data showed that ds-iniSeq origin relative activities were higher at sites where transcription took place and at sites that H3K9ac was present (Fig.5.13/14), which is consistent with the observation that H3K9ac is present at active TSS and origins. However, I found that on a genome-wide scale, there was no inverse relationship between transcription

and origin firing, as in these EJ30 cells there is no correlation between local RNA expression and ds-iniSeq origin relative activity (Fig.5.11D). As such, my data casts doubt on the Dfb4 promoter hypothesis of co-regulation.

My data showed that there was no enrichment of ds-iniSeq origins at H3K9me3 sites genome-wide (Fig.5.13). This directly contradicts the supposition that H3K9me3 is present during the early stages of DNA replication but the colocalisation of origins with H3K9me3 later during and after DNA replication cannot be ruled out, giving some merit to the Dfb4 promoter hypothesis, whereby H3K9ac/me3 co-ordinate transcription and DNA replication temporally. To address this issue, one must identify origins that fire later in DNA replication, which is something I attempt in my ds-eloSeq chapter. Another potential explanation for the differences in H3K9me3 association with origin found in the Dfb4 promoter study (68) and my ds-iniSeq origin analysis, is the impact of cell type specificity.

H3K9ac associates with origins that fire in multiple cell lines from different lineages (shared origins) but H3K9me3 associates with origins in specific cell lineages (specific origins) (34). The EJ30 and/or HCT116 cell lines (used in the ChIP-Seq analysis I utilised for my computational analysis) may be of cell lineages that do not require H3K9me3, but those human cell lines in the Dfb4 promoter study do (68). To determine if the EJ30s are a cell line that possessed an association of origins with H3K9me3, I would like to have conducted ChIP-Seq for H3K9ac and H3K9me3 in the EJ30 cell line.

In addition to cell lineage differences, H3K9me3 associates with origins differently in cancerous and non-cancerous cells of the same lineage. Du *et al* (69) showed that H3K9me3 was enriched at late replicating foci in normal prostate cells (PrEC) but its association with replication foci was evenly distributed throughout DNA replication in cancerous prostate cells (LNCaP). This demonstrated that the cancerous nature of a cell line may have a significant effect on the association of H3K9me3 with replication origins. As EJ30s are a cancer cell line, a future avenue of research may include attempting to convert the ds-iniSeq method to a non-cancerous male bladder cell line, in order to assess how substantial the differences in origin colocalisation levels with epigenetic features are between cancerous and non-cancerous bladder cell lines.

LNCaP is an epithelial prostate carcinoma cell line, which is reasonably similar to the male bladder epithelial carcinoma, EJ30s. However, both cell lines clearly showed substantial differences to one another, in H3K9me3 association with DNA replication in early S-phase, which highlights the considerable impact of cell lineage differences.

H3K27 is another histone position that has had two different histone modifications associated with different stages of DNA replication origin firing (28). Although no literature implies an antagonistic relationship between H3K27ac and H3K27me3, as seen with H3K9, both

modifications would not be present on the same histone at the same time. The ChIP-seq data supported this assumption and showed negligible overlap between H3K27ac and H3K27me3 (Fig.5.15B). The ds-iniSeq origins found to associate with H3K27ac and/or H3K27me3 also demonstrated a high degree of mutual exclusivity (Fig.5.15B).

As with H3K9, H3K27ac colocalises with early firing origins (28,34), which was supported by my data (Fig.5.13), and H3K27me3 associates with later firing origins (57). As previously discussed, my ds-iniSeq origins represent origins that fire earlier in DNA replication and their strong colocalisation with H3K27ac and poor association with H3K27me3 is consistent with their bias for early firing origins (Fig.5.13). The relative activities of the ds-iniSeq origins colocalised with H3K27ac were unsurprisingly high and in accordance with other EFOA marks, but the relative activities of the H3K27me3 overlapping ds-iniSeq origins were not as low as those overlapping another LFOA mark (Fig.5.14).

There is contradictory literature with respect to the prominence of H3K27me3 at early and late replicating origins; a substantial body of evidence found that H3K27me3 predominantly colocalises with late firing origins (57), but others showed that ~40% of early- and mid- S-phase firing origins associates H3K27me3 (5). Clearly my data supports the former.

Nevertheless, there remains a potential explanation for this contradictory evidence; namely H3K27me3 has been shown to associate with late replicating loci in LNCaP cells and with early replicating loci in PrEC cells (69). As ds-iniSeq used a cancerous cell line (EJ30), this marked difference in H3K27me3 distribution in cancerous vs non-cancerous cell lines may be applicable to these EJ30s and explain why ds-iniSeq origins showed a disinclination for H3K27me3.

Additionally, H3K27me3 is associated with initiating DNA replication in a cell-type specific fashion (34) and cell lineage differences may have been present between the EJ30 and HCT116 cells. I performed the same overlap analysis for all the histone marks in this analysis with ChIP-seq data from 3 different cell lines (HeLa S3-cervical carcinoma; H1-human embryonic stem cell; PC3-prostate adenocarcinoma) and found highly similar colocalisation levels (data not shown). This gives me greater confidence in the accuracy and authenticity of these findings.

H3K27me3 has been previously discussed, in conjunction with H3K4me3, as the two marks that define bivalent promoters, in order to regulate transcription (32–34). The influence of this bivalent state may extend to replication origin firing, particularly in later firing origins. This may be relevant to the relative activities of ds-iniSeq origins overlapping H3K27ac and/or H3K27me3. Although origins overlapping H3K27me3, possessed relative activities lower than those overlapping its acetylated counterpart, the range remained similar and did not show a greater accumulation of origins with lower relative activities (Fig.5.14&15B). This was

unlike the analysis of H3K9ac/me3 (Fig.5.15&15A). If H3K27me3 was solely a histone mark associated with late firing origins, one might expect to see a similar pattern in relative activities as H3K9me3, but this was not the case and may be explained by H3K27me3's role in the production of poised promoters/bivalent states.

As with other histone marks discussed here, H3K27me3 and H3K27ac have allocated roles in transcription; H3K27me3 is linked to transcriptional silencing and H3K27ac is a hallmark of transcriptional activity (57,70). In line with expectation of an active transcription histone mark, H3K27ac has been found at OK-seq sites associated with active TSS. Interestingly, H3K27ac has also exhibited an enrichment at enhancers (7,71). I have found negligible colocalisation of ds-iniSeq origins with enhancers (Fig.5.12D/E). Therefore, it is unlikely that the association of H3K27ac with enhancers is related to origin firing and makes a stronger argument for H3K27ac's involvement in transcription. In addition to its association with early firing origins and OK-seq sites, H3K27ac acts as a predictor for ORC binding, colocalises with MTBP-binding sites (7,28,34,49,65) and colocalises with ds-iniSeq origins that possessed high relative activities (Fig.5.14), which gives further evidence for its role in DNA replication and origin firing. These findings highlight the deeply intertwined nature of DNA replication and transcription and suggest that H3K27ac may have a role to play in the regulation of both.

The mutation of H4K20 partially impairs S-phase progression and prevents re-replication, which highlights its importance in DNA replication and origin firing (44). The PR-Set7 methyltransferase is solely responsible for the mono-methylation of H4K20. PR-Set7 and H4K20me1 have been suggested as regulating ORC recruitment to chromatin, and therefore origin licensing, which is supported by high levels of H4K20me1 during G1 but lower levels across S-phase (43,72). As S-phase progresses, PR-Set7 undergoes PCNA-Cul4-driven degradation and H4K20me1 levels reduce, presumably through the conversion of H4K20me1 to H4K20me2/3 and the absence of PR-Set7. This has been shown to inhibit further origin licensing (45,72,73). The transition of H4K20me1 to H4K20me2/3 by Suv4-20h may explain the lack of overlap between H4K20me1 ChIP-seq sites and ds-iniSeq origins (Fig.5.13). To ascertain if this were the case, I would need to obtain some ChIP-seq data for H4K20me2 and H4K20me3 (HCT116) and perform an overlap analysis with these data sets. H4K20me2/3 are believed to act as an enhancer for MCM2-7 loading and the activation of a subset of origins present in late-replicating heterochromatin domains (44). The lack of colocalisation between the ds-iniSeq origins and H4K20me1 supports the findings by Kumagai & Dunphy (49) who found no colocalisation of MTBP-binding sites with H4K20me1. They also showed no correlation between MTBP-binding and H4K20me3 sites. I would therefore expect a fairly small colocalisation between H4K20me3 and my ds-iniSeq origins but would anticipate an increase in colocalisation with origins that appear later in replication.

The ds-iniSeq origin relative activity of the H4K20me1 overlapping origins showed a moderately similar distribution as those origins associated with EFOA marks (Fig.5.13&14). This may be indicative of a lack of influence of H4K20me1 on origin activation, as H4K20me1 is considered to be a factor involved in origin licensing rather than firing.

Most replication origin histone marks are also attributed to roles in transcription. The relationship between these EFOA epigenetic features and active transcription is clearly evident; H3K27ac, H2A.Z, H3K79me2, H3K4me3 are hallmarks of transcriptional activity (28,70). This relationship is expected as both origin firing and transcription have been found to take place on the same regions of DNA. This does result in a great deal of difficulty in being able to distinguish the features responsible for each process, from one another.

The histone mark H3K36me3 has also been identified as a hallmark of transcription and there remains limited evidence of a role in DNA replication but it has been suggested that it may associated with early firing origins (7,28). Due to its association with transcriptional activity, if H3K36me3 plays a role in origin specification/activation, one would expect H3K36me3 to strongly colocalise with highly active, early firing ds-iniSeq origins.

I found no enrichment of H3K36me3 at the ds-iniSeq origins, nor did the ds-iniSeq origins overlapping H3K36me3 have exclusively high relative origin activities (Fig.5.13&14). I concluded that H3K36me3 is not a histone mark that defines replication origin firing in early replication, despite its prominent role in transcription. This is corroborated by the findings that MTBP-binding sites also did not colocalised with H3K36me3 (49).

These findings raise the possibility that some transcriptionally active epigenetic marks may have been mistakenly attributed to a role in DNA replication origin firing or to a role in transcription.

Overall, the current literature, in conjunction with my findings suggest a sophisticated, complex, and multi-layered interaction of multiple histone marks at DNA replication origins that ultimately facilitate origin firing.

The importance of CGI-promoters and their interaction with EFOA marks has been demonstrated here (Fig.5.16A-C). As stated previously, there is no known consensus sequence for human origins and the number of genome-wide CGIs and CGI-promoters vastly outweighs the number of ds-iniSeq origins (Fig.5.12C&16) (28,74,75), which would further implicate additional factors in origin specification and activation. I propose that these factors are histone marks that may be facilitated by the presence of CGI-promoters and this is supported, in part, by aspects of the literature.

CGIs/CGI-promoters act as targets for histone modifiers. CGI-promoters can regulate factors that post-transcriptionally modify histones, which then modulates gene expression (76). The

transcriptionally active, H3K4me3, is a signature mark for CGI-promoters and this enrichment of H3K4me3 at CGI-promoters correlated with increased transcriptional activity. CGI-promoters also govern H3K4me3 distribution in the genome; CGI-promoters direct complexes that promote H3K4 tri-methylation. In turn, H3K4me3 further influences the chromatin landscapes (76–78).

Additionally, CGIs/CGI-promoters are targeted by PRC2 which tri-methylates H3K27, which is a marker for transcriptional silencing. As previously discussed, H3K4me3 and H3K27me3 are two marks that define a bivalent state at promoters. In the context of transcription, this bivalent state determines the degree of transcriptional activity which is dependent upon the ratio of H3K4me3 and H3K27me3 (30,31,41,53,55,79,80). It may be possible that this bivalent relationship is applicable to DNA replication, where H3K27me3 has more influence in later firing origins.

One further feature of CGI-promoters is that they adopt a transcriptionally permissive state. This state does not determine the degree of transcriptional activity but acts as a specific signal that can initiate the transition of a permissive state to an active one. In all likelihood, this signal would be H3K4me3 as it influences the chromatin landscape (41,76,81). It is possible that H3K4me3 promotes the deposition of other EFOA histone marks, thus shifting the CGI-promoter permissive DNA into active replication, which would result in higher relative origin activities (ie greater probability of origin firing).

Furthermore, origins that are shared between different cell line lineages are strongly associated with CGIs, in addition to H3K4me3 and H3K9ac (34). Overall, I propose that CGI-promoters and CGIs directs the tri-methylation of H3K4, which in turn drives the generation of additional EFOA histone marks, such as H3K9ac and ultimately bring about a chromatin environment that is conducive with origin activation/firing. Further work is now needed to establish the nature of relationships between CGIs and CGI-promoters with EFOA histone marks, and their subsequent potential causal relationship with origin specification and activation.

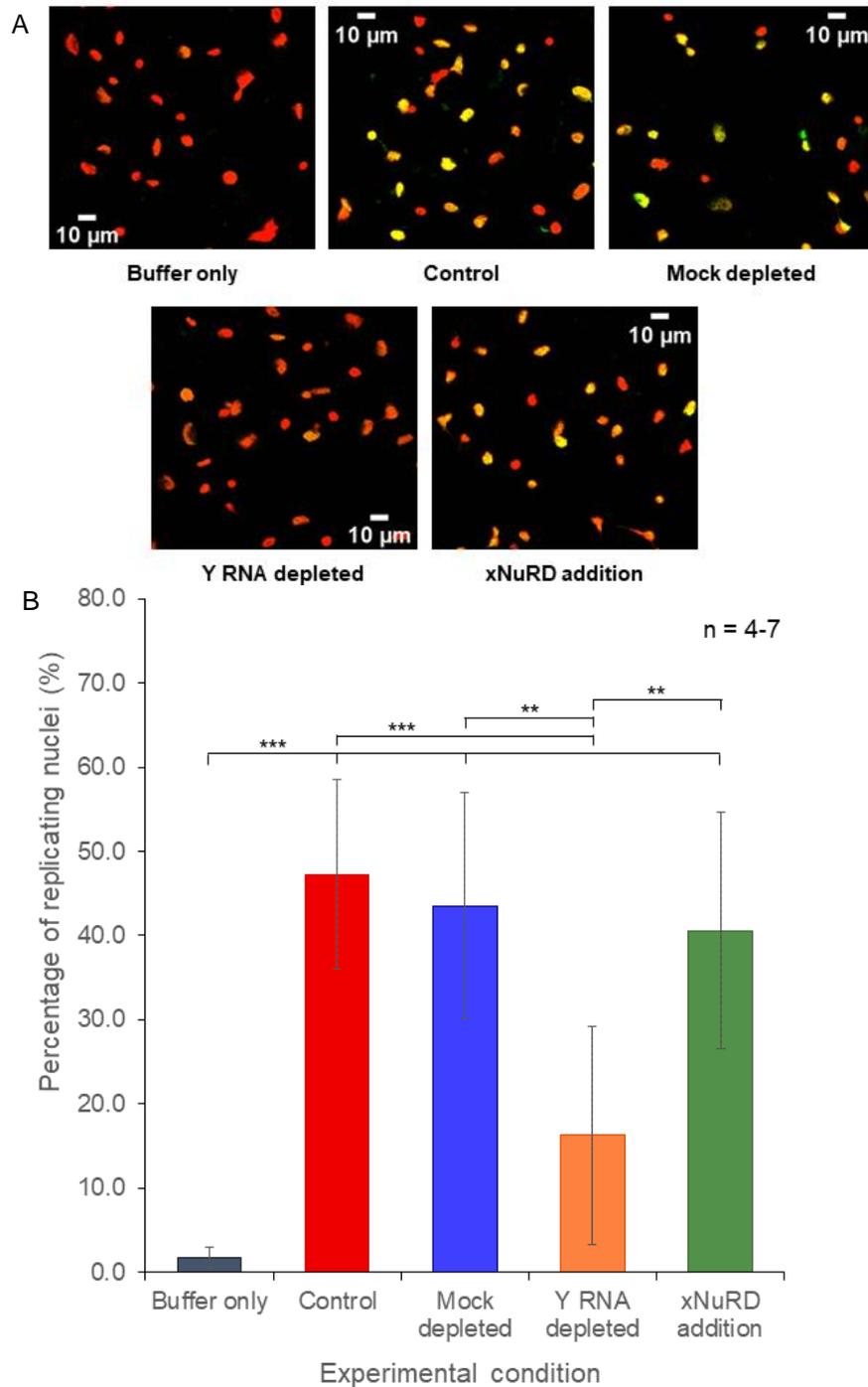
Overall, in this chapter, I have shown an extensive analysis of ds-iniSeq origins with alternative NGS methods for origin identification, and genomic and epigenetic features associated with replication origins. In particular, I have identified CGIs and CGI-promoters as features associated with highly active replication origins and then demonstrated a link between CGIs/CGI-promoters, EFOA histone marks and ds-iniSeq origin relative activity. These have provided a solid foundation for further analysis. I can manipulate the replication reactions in the ds-iniSeq protocol to assess the effect of essential DNA replication factors, with the aim to help further elucidate the mechanisms by which human DNA replication origins are specified and activated.

## **Chapter 6: The role of hY RNAs and xNuRD for the activation of human DNA replication origins, as determined by ds-iniSeq**

### **6.1 Introduction**

Y RNAs have been identified as essential DNA replication factors and the nucleosome remodeller xNuRD has been identified as a Y RNA-independent DNA replication initiation factor. Importantly, xNuRD can functionally substitute for Y RNAs in the human cell-free system. The mechanism by which these factors carry out their roles remains unelucidated (1,2). In this chapter, I will demonstrate the importance of Y RNAs and xNuRD and investigate their role in DNA replication origin activation.

Figure 6.1 shows an example of the effect of Y RNA removal and replacement with xNuRD on DNA replication in the human cell-free system.



**Figure 6.1:** The human cell-free experiments that demonstrate the effect of Y RNA removal and xNuRD addition. Template nuclei, isolated from mimosine treated EJ30 cells (synchronised to late G1), were incubated (3 hours; 37°C) in the presence of a physiological replication buffer, NTPs, dNTPs (including dig-dUTP), creatine kinase and in the absence of cytosol (“Buffer only”); with cytosol (proliferating HeLa cells) (“Control”); with cytosol (proliferating HeLa cells) pre-incubated with an oligonucleotide complementary for the bacteriophage T3 RNA, to act as a control for oligonucleotide/RNase H degradation of RNAs (“Mock depleted”); with cytosol (proliferating HeLa cells) pre-incubated with an oligonucleotide complementary for hY RNAs, to degrade/eliminate hY RNAs (“Y RNA depletion”); and with cytosol (proliferating HeLa cells) pre-incubated with an oligonucleotide complementary for hY RNAs, to degrade/eliminate hY RNAs, and the addition of partially purified xNuRD (“xNuRD addition”). Following the replication reaction incubation, the nuclei for each condition were transferred to individual microscope slide and stained. The nuclei were imaged using confocal microscopy and example images for each condition are shown in (A). (B) From these confocal images, the nuclei were counted to provide percentages of replicating nuclei (n = 4-7). An ANOVA and subsequent Tukey’s post-hoc test were performed to assess significance; the Tukey’s test results are shown on the plot and \*\*\* indicates  $p < 0.001$  and \*\* indicates  $p < 0.01$ ; all other tests were not significant

Confocal images (Fig.6.1A) showed examples of the impact of Y RNA removal and xNuRD addition on DNA replication in the isolated template nuclei. Synchronised late G1 nuclei were incubated for 3 hours, with a physiological replication buffer, NTPs, dNTPs (including dig-dUTP), an ATP-regeneration system, and either no cytosol, untreated cytosol (from proliferating HeLa cells), mock-depleted cytosol (incubated cytosol with non-human specific DNA oligonucleotide), Y RNA-depleted cytosol (incubated cytosol with DNA oligonucleotides specific for hY RNAs) or Y RNA-depleted cytosol with xNuRD. The nuclei were transferred to slides and stained with propidium iodide for DNA (red) and fluorescently tagged anti-dig antibodies for dig-dUTP incorporation in newly synthesised DNA (green). In these images, non-replicating nuclei appeared red and replicating nuclei appeared yellow/orange, resulting from the dig-dUTP incorporation.

The buffer only control (no cytosol) had mostly red nuclei, indicating no DNA synthesis/replication in the vast majority of template nuclei. The control (untreated cytosol) showed a high degree of dig-dUTP incorporation into a large proportion of nuclei. The mock-depleted condition showed similar levels of dig-dUTP incorporation/DNA replication as the control. The Y RNA-depleted condition showed a marked decrease in dig-dUTP incorporation/DNA replication compared to the control and mock-depleted conditions. The addition of xNuRD to the Y RNA-depleted cytosol (xNuRD addition), increased dig-dUTP incorporation/DNA replication compared to the Y RNA-depleted control, with DNA replication levels returning to similar levels to that of the control and mock-depleted conditions.

The nuclei detected in the confocal images were categorised as non-replicating (red) and replicating (yellow/orange) nuclei and counted to determine the percentage of replicating nuclei (Fig.6.2). The percentages of replicating nuclei reflected the pattern observed in the confocal images. The buffer only control possessed very few replicating nuclei (1.7%), which indicated the degree of contamination by S-phase nuclei. S-phase nuclei were the only nuclei able to replicate in the absence of cytosol, as replication initiation had already begun *in vivo*; whereas G1-phase nuclei required the essential replication factors present in cytosol to initiate DNA replication.

Addition of cytosol or mock-depleted cytosol to the replication reaction resulted in an increase of replicating nuclei (47.2% & 43.5% respectively), indicating that >40% of G1-phase nuclei had initiated DNA replication *in vitro*, which is consistent with published literature (1,2). When Y RNAs were removed, the percentage of replicating nuclei decreased to 16.3%, demonstrating the essential nature of Y RNAs in DNA replication initiation. This significant decrease in DNA replication did not return levels to the same as the buffer only control. This may reflect the effect of residual Y RNAs that had escaped RNase-H degradation. When xNuRD was added to the Y RNA-depleted cytosol, the replication nuclei percentage returned to levels (40.6%) similar to the control and mock-depleted conditions,

highlighting the established role of xNuRD as a Y RNA-independent DNA replication initiation factor (2).

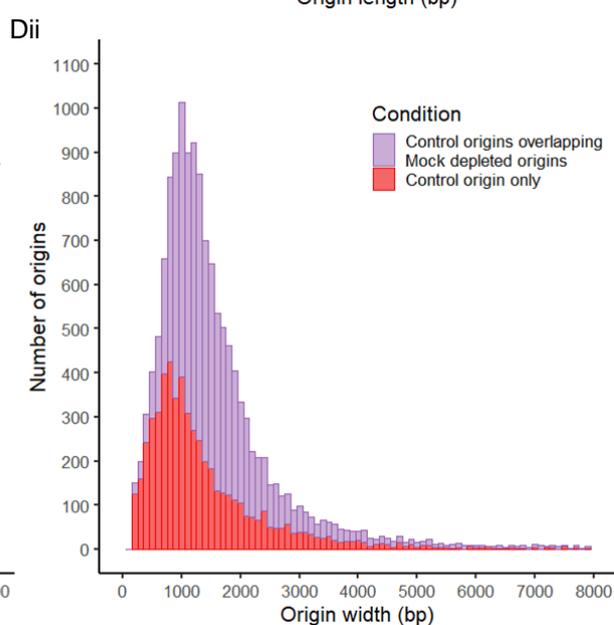
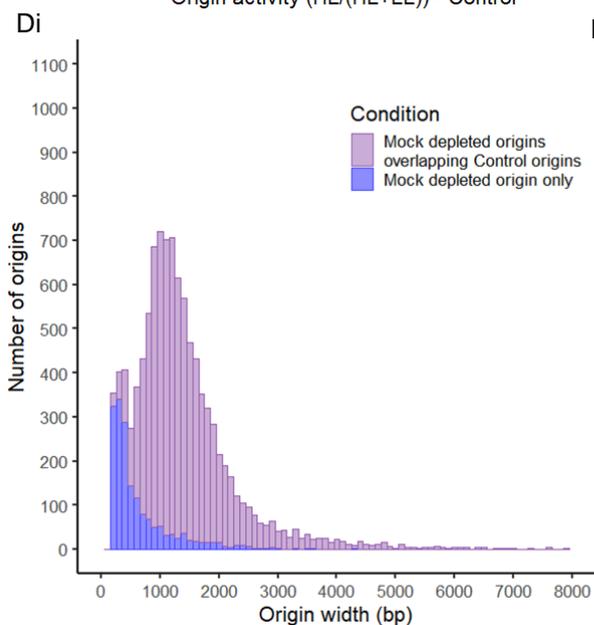
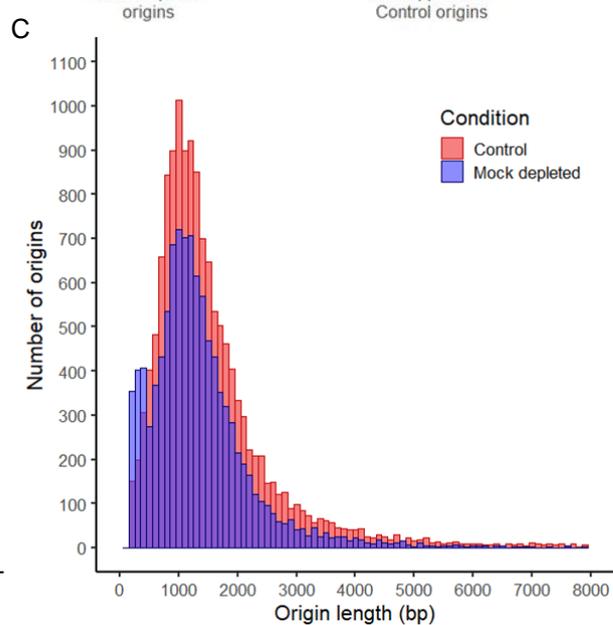
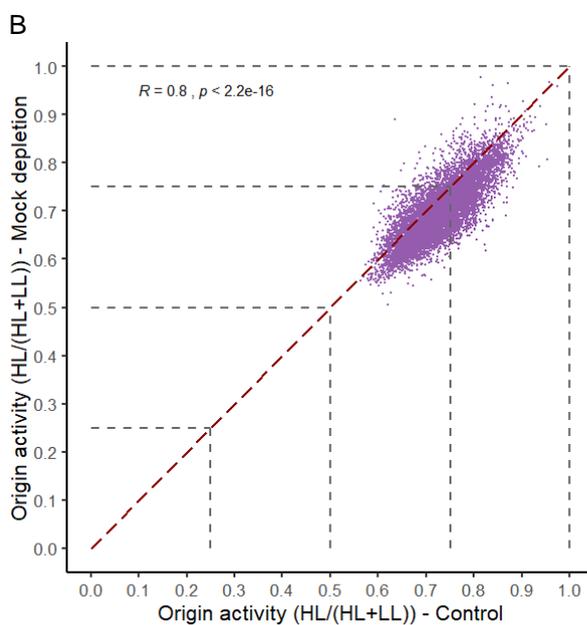
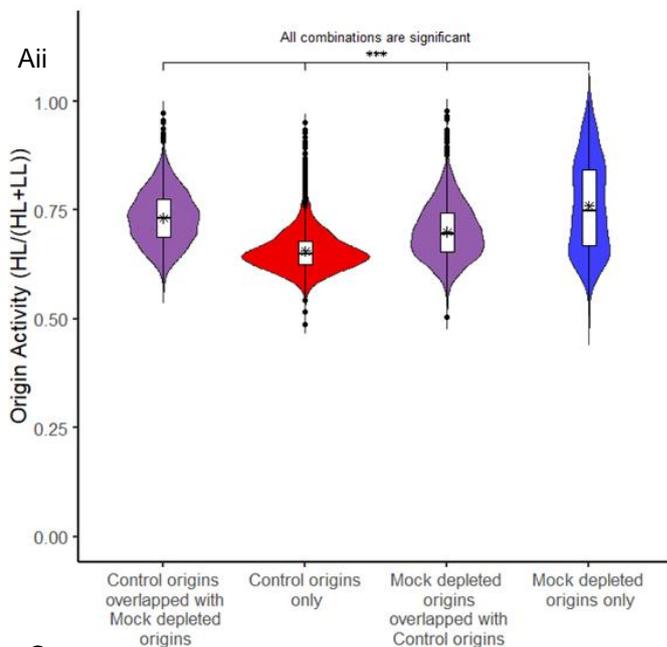
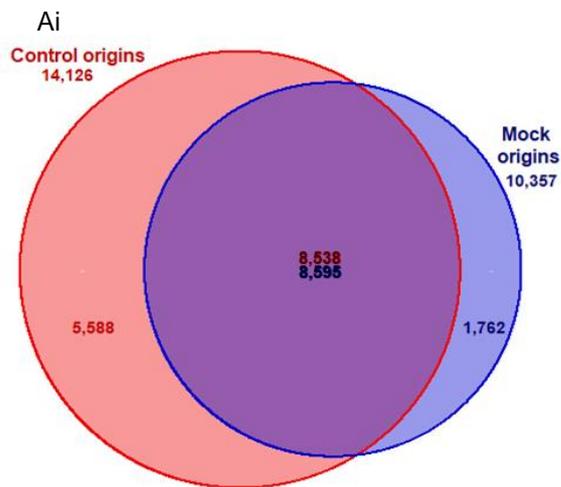
I have confirmed the essential nature of hY RNAs and xNuRD on DNA replication initiation in human G1-phase nuclei, as published (1,2). hY RNAs act in human cell nuclei with discrete origin sites which become activated during replication initiation; similar is seen with their homologue xY RNAs in *X. laevis* (3,4). However, xNuRD was identified in pre-MBT *X. laevis* embryos where delocalised DNA replication and rapid and overlapping cell cycles take place (3,5,6). Firstly, the mechanism by which Y RNAs play their essential role in human DNA replication initiation is currently unknown. Secondly, the mechanism by which xNuRD fulfils its role in DNA replication and whether it performs the same function as Y RNAs also remains unclear. Furthermore, the effect of these factors on the specification and activation of human DNA replication origins remain unelucidated. In this chapter, I aim to address the two key questions; How do Y RNAs affect human replication origin specification and activation? And how does xNuRD affect human replication origin specification and activation in the absence of Y RNAs? To achieve this, I performed analyses using ds-iniSeq experiments where Y RNAs were removed and xNuRD added.

## 6.2 Results

I performed three replicates of the ds-iniSeq experiments with the mock-depleted, Y RNA-depleted and xNuRD addition conditions. As detailed for control ds-iniSeq experiment (discussed in chapter 5), I generated a list of ds-iniSeq origins present in replicate 1 that were also found in replicates 2 & 3 for each condition, which were used for all subsequent analyses in this chapter. Those origins identified by ds-iniSeq under the control conditions (ie standard ds-iniSeq reaction – see chapter 5) will be referred to as control origins. Origins identified by ds-iniSeq under mock depletion, Y RNA depletion and xNuRD addition (following Y RNA depletion) will be referred to as mock-depleted origins, Y RNA-depleted origins and xNuRD addition origins, respectively.

### 6.2.1 Comparison of Control ds-iniSeq origins and Mock-depleted ds-iniSeq origins

Initially I compared the control origins with the mock-depleted origins; the mock-depleted condition acted as the positive control for all later analysis. I performed an overlap analysis and compared relative origin activities and widths (Fig.6.2).



**Figure 6.2:** The comparison of the control ds-iniSeq origins and the mock-depleted ds-iniSeq origins. (Ai) The degree of overlap between control ds-iniSeq origins (red) and mock-depleted ds-iniSeq origins (blue). (Aii) The origin activity of the control ds-iniSeq origins that do (purple (left)) and do not (red (left)) overlap with mock-depleted ds-iniSeq origins; and the origin activity of the mock-depleted ds-iniSeq origins that do (purple (right)) and do not (blue (right)) overlap control ds-iniSeq origins. The mean is indicated with an \* and an ANOVA and subsequent Tukey's post-hoc test were performed to assess significance; the Tukey's test results are shown on the plot and \*\*\* indicates  $p < 0.001$ . (All test combinations were significant to  $p < 0.001$ ). (B) The scatter of the origin activities of the control (x axis) and mock-depleted (y axis) ds-iniSeq origins that overlapped one another. A linear regression, Pearson test for correlation and ANOVA test for significance were conducted. The correlational R value and the ANOVA test result are indicated on the plot. The dark red dashed line showed the perfect correlation ( $R = 1$ ) for comparative purposes. (C) The histogram of the frequency of widths of the control (red) and mock-depleted (blue; overlaying the control widths) ds-iniSeq origins, from 0 – 8Kb/8000bp in width. (Di) The histogram of the frequency of widths of the mock-depleted origins that did (purple) and did not overlap control origins (blue). (Dii) The histogram of the frequency of widths of the control origins that did (purple) and did not overlap mock-depleted origins (red). (bin width = 100bp)

The number of ds-iniSeq origins reduced from 14,126 (control) to 10,357 when the cytosol was subjected to mock depletion. There was a high degree of overlap between the two conditions, where 60.4% control origins overlapped with 83.0% mock-depleted origins (Fig.6.2Ai).

Control origins that overlapped the mock-depleted origins were significantly more active than control origins that did not overlap with mock-depleted origins (control origins only) (Fig.6.2Aii). The distribution of activities of the control only origins (non-overlapping) also showed an accumulation of lower activities, revealing that origins common to both conditions possessed higher origin activities and thus a higher probability of firing/initiating.

The activities of the mock-depleted origins that overlapped with the control origins, possessed similar activities than the control origins overlapped with mock-depleted origins. This implies that those origins present in both conditions retained similar probabilities for firing/origin activities (Fig.6.2Aii).

Furthermore, I found a very strong positive correlation ( $R=0.8$ ) between the relative activities of the ds-iniSeq origins common to the mock-depleted and control conditions (Fig.6.2B), which is further evidence for the above deduction.

By contrast, mock-depleted origins that did not overlap with the control origins (17%; Fig.6.2Ai) possessed the significantly highest overall origin activities, with the largest range of activities (Fig.6.2Aii). On average, they had a higher probability of firing/initiating than those that did overlap, suggesting that mock depletion treatment results in the effective activation of a few additional origins.

Comparison of the widths of the control and mock-depleted origins (Fig.6.2C) showed that both conditions had a similar peak at widths at ~1.2Kb. Origin widths indicated the extent of DNA replication at these origins. Generally, the mock-depleted origin widths followed the same distribution of widths, merely with fewer origins overall. However, the mock-depleted

ds-iniSeq origins had an initial “sub-peak” with an accumulation of much shorter origins at 200-500bps; whereas the control origins did not (only one single large peak at ~1.2Kb).

To further investigate the curious nature of the highly active mock-depleted origins that did not overlap with the control origins, I evaluated the origin widths of the mock-depleted only origins and the mock-depleted origins overlapping control origins (Fig.6.2Di). Those mock-depleted only origins that possessed high origin activities (Fig.6.2Aii) were highly skewed to lower widths and constituted the bulk of the initial “sub-peak” at 200-500bps, whereas the mock-depleted origins that overlapped the control origins possessed widths similar to those of the overall control origin widths (peak ~1.3Kb).

For comparison, I performed the same assessment on the control origins (Fig.6.2Dii). The control only origins were also skewed to the shorter widths, but the distribution was much wider, and the peak was larger (~1.1Kb), than the mock-depleted only origins.

The widths of the control origins overlapping the mock-depleted origins (Fig.6.2Dii) showed an almost identical distribution to those mock-depleted origins overlapping the control origins (Fig.6.2Di). They both had a large single peak of origin widths around ~1.4Kb (control) and ~1.3Kb (mock-depleted), with very few origins extending to widths of 39.3Kb (control) and 22.3Kb (mock-depleted). Overall, there were more control origins at each width group than its mock-depleted counterpart. As with the comparison of activities of the ds-iniSeq origins present in both experimental conditions, this appeared to demonstrate that these origins possessed a similar extent of DNA replication. However, the control origins experienced DNA replication that had progressed slightly further.

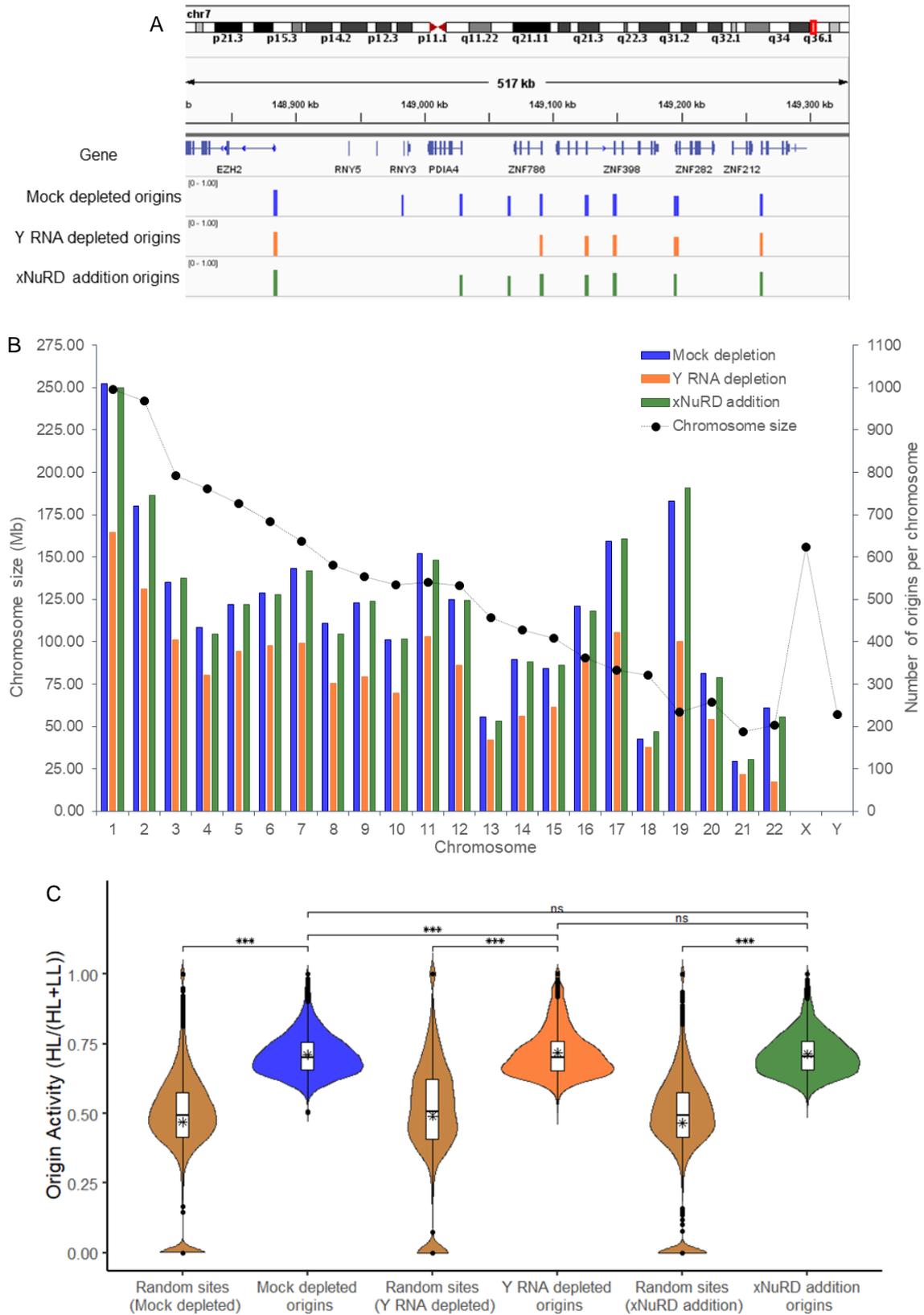
Taken together, this data suggests that mock depletion has a mild global impact on the extent of DNA replication of most of ds-iniSeq origins. Most origins shared genomic location and relative initiation activities between these conditions. Mock depletion also activated a small number of new small but highly active ds-iniSeq origins.

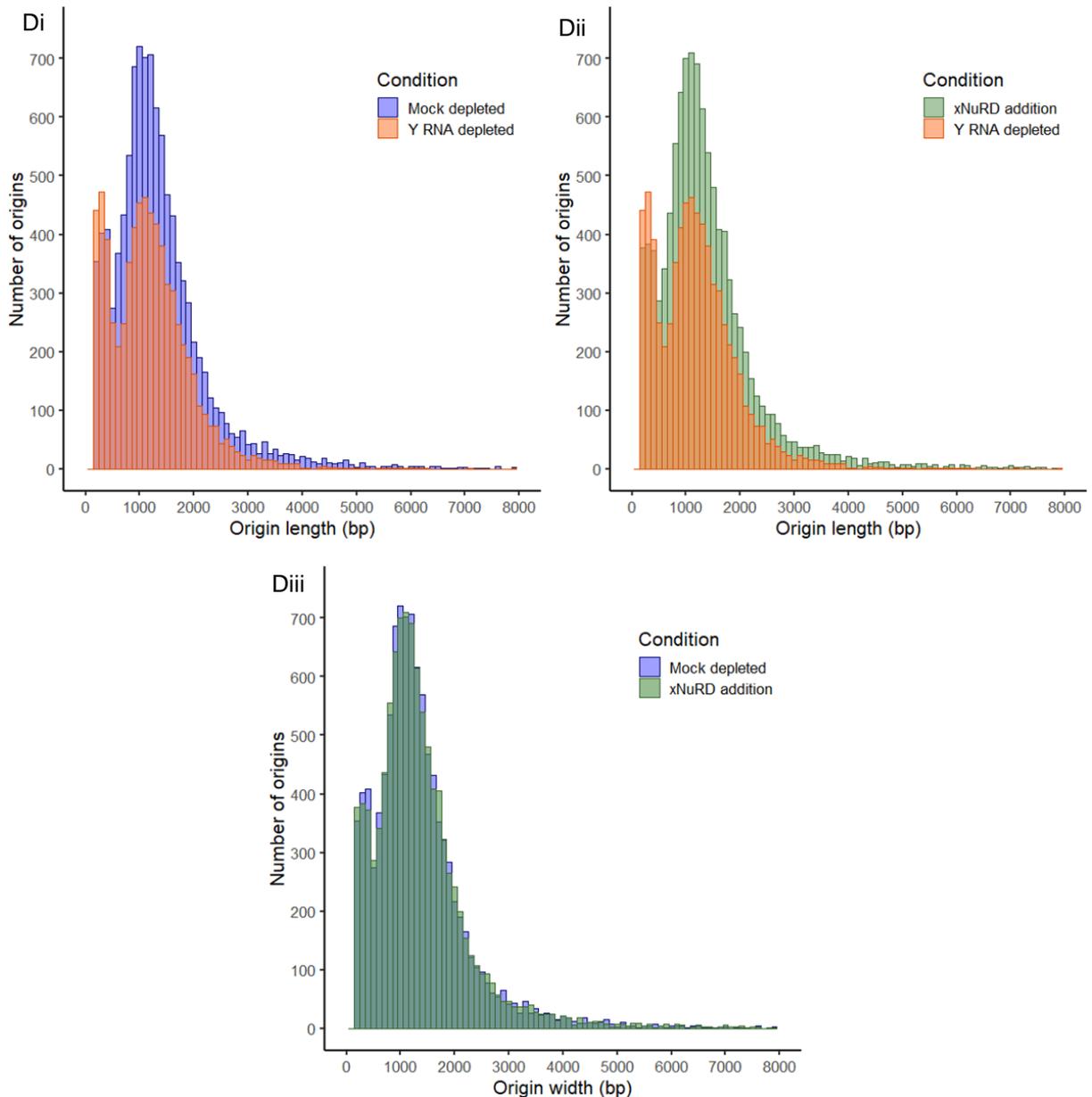
From these data (Fig.6.2), it appeared that the addition of the unspecific DNA oligonucleotide (mock-depleted) had a modest effect on replication origin firing. I therefore adopted the mock-depleted origins as a control for the rest of the analysis during this chapter, because the Y RNAs were depleted in the subsequent reactions using specific DNA oligonucleotides, complementary to the hY RNAs.

### *6.2.2 Comparison of Mock-depleted, Y RNA-depleted and xNuRD addition ds-iniSeq origins*

I compared the mock-depleted origins to the ds-iniSeq origins identified in the absence of Y RNAs (Y RNA-depleted) and the presence of xNuRD (xNuRD addition). There were 10,357 mock-depleted origins, which reduced to 7,076 origins upon Y RNA removal and then increased back up to 10,329 origins when xNuRD was subsequently added. This indicates that Y RNA depletion results in the inhibition of DNA replication initiation at the level of activated origin numbers and the addition of xNuRD leads to the emergence of new active origins. The origins in the Y RNA-depleted and xNuRD addition conditions must be further interrogated to establish if Y RNAs and xNuRD target the same origins.

I assessed the effect of these experimental conditions on the number of origins per chromosome, relative origin activities and origins widths (Fig.6.3).





**Figure 6.3:** (A) An IGV image of the mock-depleted (blue), Y RNA-depleted (orange) and xNuRD addition (green) ds-iniSeq origins, at the region near the EZH2 gene. (B) The number of origins per chromosome (secondary Y axis) for mock-depleted (blue), Y RNA-depleted (orange) and xNuRD addition (green) ds-iniSeq origins were compared to the size (Mb) of each chromosome (black scatter on primary Y axis). (C) The origin activities of the mock-depleted (blue), Y RNA-depleted (orange) and xNuRD addition (green) ds-iniSeq origins compared to their corresponding random sites (ochre). The mean is indicated with an \* and an ANOVA and subsequent Tukey's post-hoc test were performed to assess significance; the Tukey's test results are shown on the plot and \*\*\* indicates  $p < 0.001$  and "ns" indicates not significant. (Di) The histogram of the frequency of widths of the mock-depleted (blue) and Y RNA-depleted (orange; overlaying the control widths) ds-iniSeq origins, from 0 – 8Kb/8000bp in width. (Dii) The histogram of the frequency of widths of the xNuRD addition (green) and Y RNA-depleted (orange; overlaying the control widths) ds-iniSeq origins, from 0 – 8Kb/8000bp in width. (Diii) The histogram of the frequency of widths of the mock-depleted (blue) and xNuRD addition (green; overlaying the control widths) ds-iniSeq origins, from 0 – 8Kb/8000bp in width. (D) bin

The IGV image of the mock-depleted, Y RNA-depleted and xNuRD addition origins (Fig.6.3A) shows a representative example of these origins. There were multiple discrete origins common to all three experimental conditions. It includes an example of an origin unique to the mock-depleted condition and two origins that were affected by/disappeared upon Y RNAs removal/depletion and then reappeared in the presence of xNuRD.

The numbers of origins per chromosome of the mock-depleted, Y RNA-depleted and xNuRD addition origins (Fig.6.3B) followed a similar pattern. The mock-depleted and xNuRD addition origins had almost identical numbers of origins per chromosome, whereas the Y RNA-depleted origins possessed the same distribution but had lower numbers than the other conditions on every individual chromosome. As with the control origins (Fig.5.1A), chromosome size appeared to have no impact on origin number per chromosome. I conclude that Y RNA removal and subsequent xNuRD addition did not disproportionately affect the number of origins on each chromosome.

The activities of the mock-depleted, Y RNA-depleted and xNuRD addition origins and their corresponding randomised sites showed that in all experimental conditions, the random sites exhibited a wide range of relative site activities (Fig.6.3C). These random site activities were significantly lower than their corresponding ds-iniSeq origins. The activities of the mock-depleted and xNuRD addition origins possessed extremely similar distributions and were not significantly different. Interestingly, the activities of the Y RNA-depleted origins were marginally greater than the mock-depleted and xNuRD addition origins. This difference was significantly greater than the mock-depleted origins but not significantly greater than the xNuRD addition origins. This significance was probably due to the distribution of the Y RNA origin activities, which showed a narrowing of the higher activities. These data indicated that overall, the origins detected in all conditions had equal probabilities for firing/initiation. This was unexpected, as the effect of Y RNA depletion on DNA replication foci in the human cell-free system was a significantly substantial reduction in incorporation of dig-dUTP and an order of magnitude lower proportion of nuclei replicating (Fig.6.1).

This may have resulted from a difference in resolution of the human cell-free system, which detects replication foci and ds-iniSeq which identifies and determines origin activation. Replication foci accumulate replicated DNA 0.5-1Mb in length and consist of multiple firing origins and subsequent elongation (7), but I demonstrated that ds-iniSeq can measure only origins to sizes below 1Kb (Fig.6.2D). The lack of consensus on the effect of Y RNAs on replication between the two methods may indicate that Y RNAs influence DNA replication following origin firing.

The number of Y RNA-depleted origins reduced when compared to the numbers of mock-depleted and xNuRD addition origins, indicating that Y RNA affected origin firing.

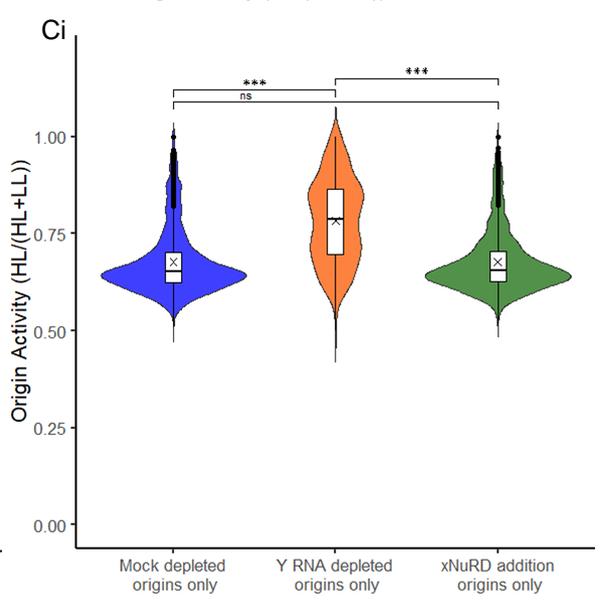
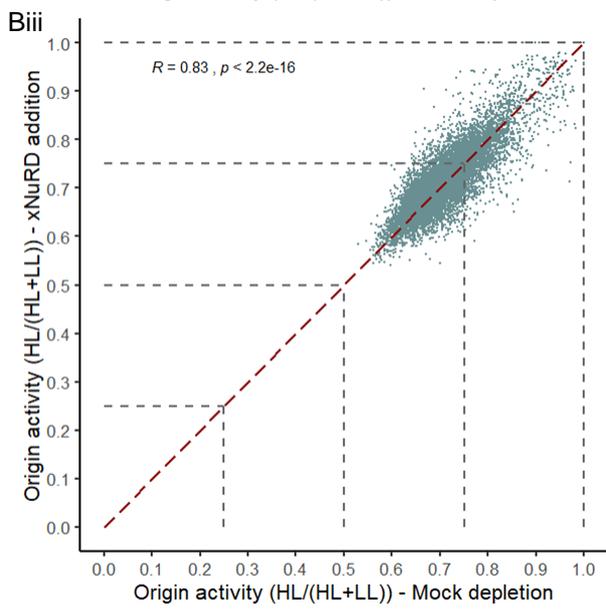
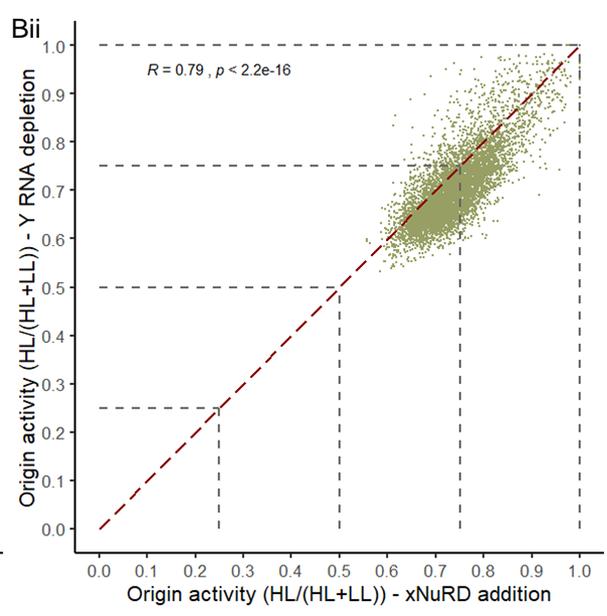
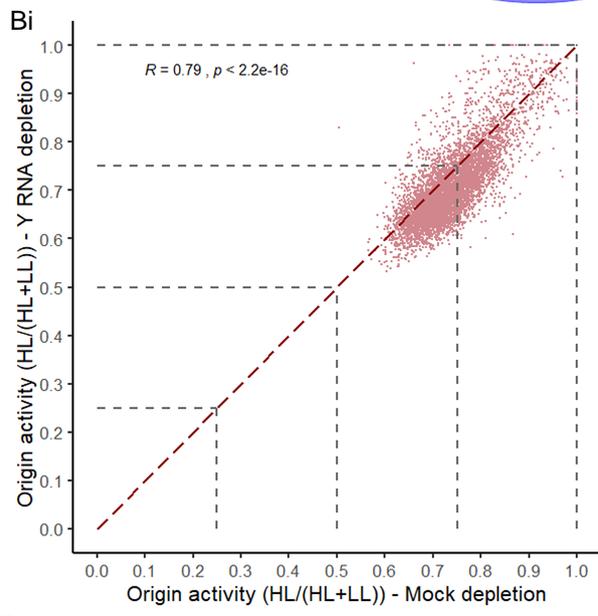
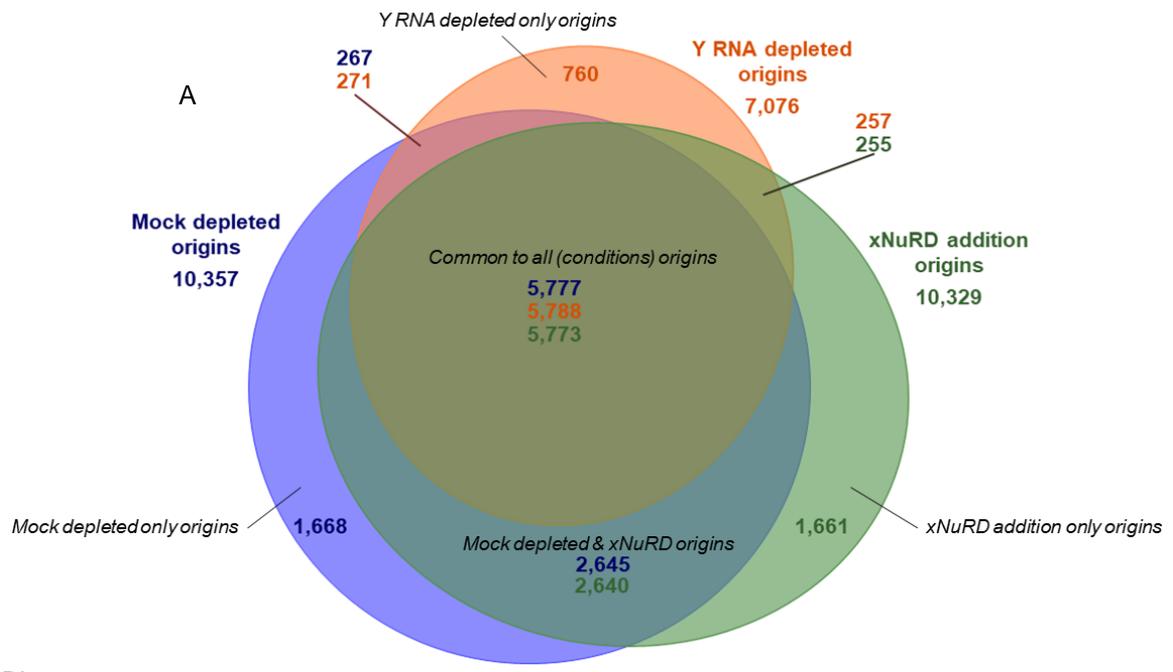
Additionally, the distribution of Y RNA-depleted origin activities was different to that of the mock-depleted and xNuRD origins, whereby the Y RNA-depleted origins possessed a flat base at lower activities and a narrow extension of activities into the higher activities (Fig.6.3C). This may be indicative of a subset of Y RNA-depleted origins that possess higher activities and skew the mean activity to higher values.

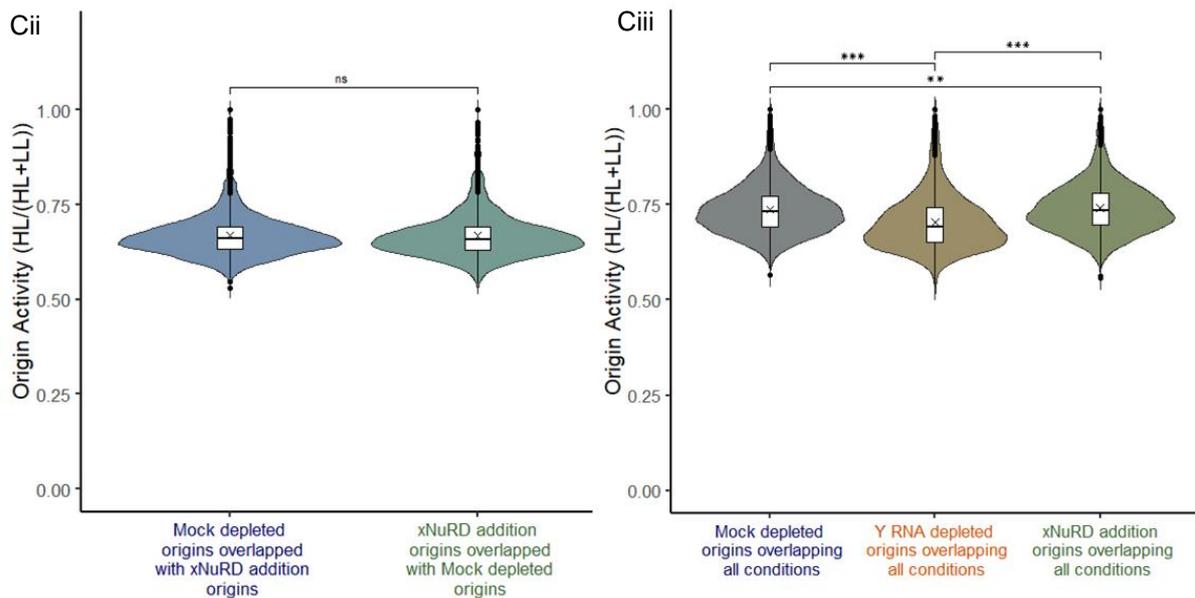
Finally, I assessed the impact of the Y RNA depletion and xNuRD addition on the extent of DNA replication via the assessment of origin widths (Fig.6.3D). The histograms showed that all of the ds-iniSeq origins possessed the same distribution of origins widths; there was an initial “sub-peak” at smaller origin widths (~200-500bps) with a second larger peak at ~1.1-1.2Kb, and the majority of widths were <3Kb. The distribution and numbers of origins for each width group of the mock-depleted and xNuRD addition origins were almost identical (Fig.6.3Diii). They both had more active of origins (~700) at widths around the peak of 1.2Kb with approximately half the number of origins (~400) at the “sub-peak” of smaller origins (~200-500bps) and the widths extended up to 22.3Kb (Mock) and 23.3Kb (xNuRD). By contrast, the Y RNA-depleted ds-iniSeq origins possessed two peaks with approximately equal numbers of origins (~450) at widths of ~200-500bps and 1.1Kb (Fig.6.3Di&Dii). The widths extended up to 12.4Kb.

These data showed that the number of origins is reduced at all widths above 400bp upon Y RNA removal, when compared to the mock-depleted origins. Upon xNuRD addition, the numbers of origins at widths above 400bp returned to similar levels as the mock-depleted origins. This suggests that the reduction in number of origins upon Y RNA removal was ostensibly distributed in an unbiased fashion across all origin widths, except for the very small origins that were not affected. Furthermore, there was a greater extent of DNA replication of both the mock-depleted and xNuRD addition origins when compared to the Y RNA-depleted origins.

### *6.2.3 How do Y RNAs and xNuRD impact DNA replication origins? – overlap and relative origin activities*

To assess whether Y RNAs and xNuRD affect the same or different origins, I performed an overlap analysis and evaluated the relative origin activities of the mock-depleted, Y RNA-depleted and xNuRD addition origins (Fig.6.4).





**Figure 6.4:** The 3 way overlap of the mock-depleted (blue), Y RNA-depleted (orange) and xNuRD addition (green) ds-iniSeq origins. The number of: mock-depleted ds-iniSeq origins present in each overlap grouping are indicated in dark blue; Y RNA-depleted ds-iniSeq origins present in each overlap grouping are indicated in dark orange; xNuRD addition ds-iniSeq origins present in each overlap grouping are indicated in dark green. (B) The scatter of the ds-iniSeq origin activities that overlapped one another in the following combinations: (i) mock-depleted (x axis) vs Y RNA-depleted (y axis); (ii) xNuRD addition (x axis) vs Y RNA-depleted (y axis); and (iii) mock-depleted (x axis) vs xNuRD addition (y axis). Linear regressions, Pearson’s test for correlation and ANOVA tests for significance were conducted. The correlational R value and the ANOVA test result are indicated on each plot. The dark red dashed line showed the perfect correlation ( $R = 1$ ) for comparative purposes. (Ci) The origin activities of the ds-iniSeq origins in the mock-depleted only overlap grouping (blue), the Y RNA-depleted only overlap grouping (orange) and the xNuRD addition only overlap grouping (green). (Cii) The origin activities of the ds-iniSeq origins present in the common to mock-depleted & xNuRD addition overlap grouping. The mock-depleted ds-iniSeq origins present in this overlap grouping are indicated on the left (blue teal). The xNuRD addition ds-iniSeq origins present in this overlap grouping are indicated on the right (green teal). (Ciii) The origin activities of the ds-iniSeq origins present in the common to mock-depleted, Y RNA-depleted and xNuRD addition overlap grouping. The mock-depleted ds-iniSeq origins present in this overlap grouping are indicated on the left (blue olive). The Y RNA-depleted ds-iniSeq origins present in this overlap grouping are indicated in the middle (orange olive). The xNuRD addition ds-iniSeq origins present in this overlap grouping are indicated on the right (green olive). The mean is indicated with an “X” and an ANOVA and subsequent Tukey’s post-hoc test were performed to assess significance; the Tukey’s test results are shown on the plot and \*\*\* indicates  $p < 0.001$ , \*\* indicates  $p < 0.01$  and “ns” indicates not significant.

I identified 10,357 mock-depleted, 7,076 Y RNA-depleted and 10,329 xNuRD addition ds-iniSeq origins. This suggested that ~3,000 origins were affected by the removal of Y RNAs. The overlap analysis of the mock-depleted, Y RNA-depleted and xNuRD addition ds-iniSeq origins showed that the bulk of these origins were present in all three experimental conditions (Fig.6.4A); 5,777 mock-depleted origins overlapped with 5,788 Y RNA-depleted origins and 5,773 xNuRD addition origins. This overlap group constituted those origins that were “unaffected” by Y RNA depletion and xNuRD addition.

Those ds-iniSeq origins that disappeared with the depletion of Y RNAs but were rescued upon the addition of xNuRD were represented by those mock-depleted origins (2,645) that

overlapped with the xNuRD addition origins (2,640). This overlap group constituted those origins that were “rescued” by xNuRD addition.

There were 267 mock-depleted origins that overlapped with 271 Y RNA-depleted origins and 257 Y RNA-depleted origins that overlapped with 255 xNuRD addition origins. There were some ds-iniSeq origins that were unique to the mock-depleted (1,668), Y RNA-depleted (760) and xNuRD addition (1,661) experimental conditions. The 1,668 mock-depleted only ds-iniSeq origins represented those origins that were affected by Y RNA depletion but were not rescued by xNuRD addition. The 1,661 xNuRD addition only ds-iniSeq origins represented the origins that only initiated in the presence of xNuRD and the 760 Y RNA-depleted only ds-iniSeq origins represented the origins that appeared in the absence of both Y RNAs and xNuRD.

I addressed the difference in relative origin activities of the mock-depleted, Y RNA-depleted and xNuRD addition origins. Initially I compared the activities of the: mock-depleted origins that overlapped with Y RNA-depleted origins (Fig.6.4Bi); the xNuRD addition origins that overlapped with Y RNA-depleted origins (Fig.6.4Bii); and mock-depleted origins that overlapped with the xNuRD addition origins (Fig.4.Biii). All three pairs of correlational analyses showed strong positive correlations, but the origins common to both mock-depleted origins and xNuRD addition conditions possessed a greater correlation ( $R\ 0.83$ ) than the other conditions ( $R\ 0.79$ )

These correlations indicated that the relative activity of any given origin was preserved in every experimental condition; origins with high activities in the mock-depleted condition were also high in the Y RNA-depleted and xNuRD addition conditions, suggesting the probability of origin firing was consistent. However, all the origins in the Y RNA-depleted condition showed a global shift towards lower activities, whilst maintaining the proportions of origins with different activities. This correlation suggests that Y RNA depletion reduced relative origin activity generally, while xNuRD addition restored this to the mock-depleted levels.

For further in-depth analysis, I assessed the effect of Y RNA depletion and xNuRD addition on the relative origin activities of those ds-iniSeq origins present in the following overlap grouping conditions: mock-depleted only, Y RNA-depleted only, xNuRD addition only, mock-depleted & xNuRD addition, and mock & Y RNA-depleted & xNuRD addition (common to all).

I assessed the activities of ds-iniSeq origins that were unique to mock-depleted (mock-depleted only), Y RNA-depleted (Y RNA-depleted only) or xNuRD addition (xNuRD addition only) experimental conditions (Fig.6.4Ci) and found that the activities of the mock-depleted only and xNuRD addition only origins were not significantly different. They possessed almost identical distributions of relative origin activities, with an accumulation of ds-iniSeq origins at lower activities. Conversely, relative activities of the Y RNA-depleted only origins were

significantly higher and showed a larger range than the mock-depleted and xNuRD addition only origins. This suggests the presence of a small subset of efficient Y RNA independent origins.

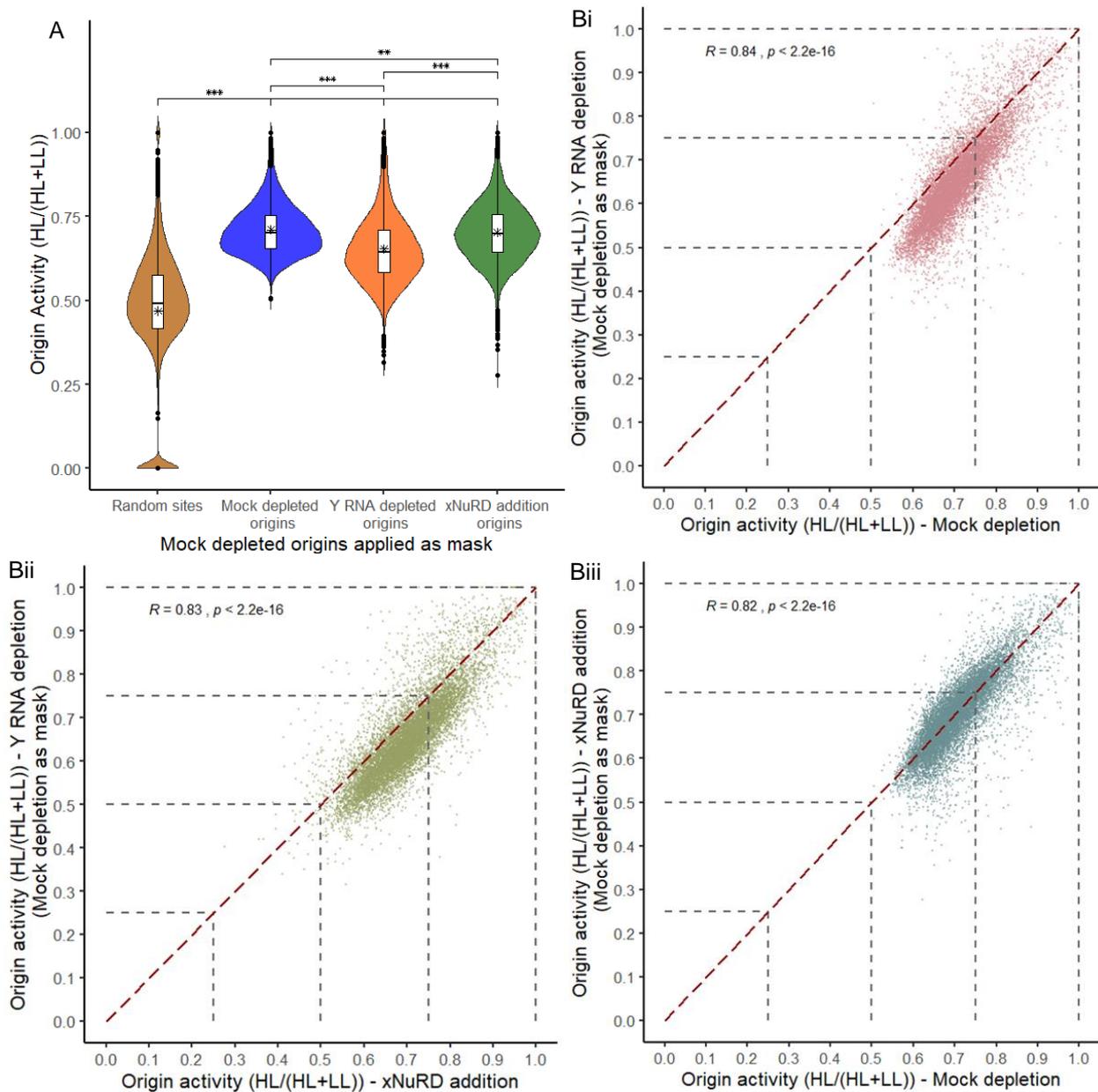
The mock-depleted & xNuRD addition overlap grouping consisted of origins common to only mock-depleted and xNuRD addition conditions and they possessed statistically similar origin activities, with large accumulations of origins at the lower relative activities (Fig.6.4Cii). This showed that those ds-iniSeq origins that were affected/disappeared by Y RNA removal but were rescued by xNuRD possessed very low relative origin activities and thus, low probabilities of firing/initiating.

Finally, I compared the relative activities of those origins present in all 3 experimental conditions (Fig.6.4Ciii). The activities of the mock-depleted origins common to all were extremely similar to that of the corresponding xNuRD addition origins; they were highly active and fairly evenly distributed. Despite their similarities, they were significantly different ( $p = 0.0075$ ), which was probably due to the minor differences in the distribution of relative origin activities.

The relative activities of the Y RNA-depleted origins common to all conditions were significantly lower than the corresponding mock-depleted & xNuRD addition origins. The distribution of the relative activities also shifted downwards, when compared to the corresponding mock-depleted and xNuRD addition origins.

These data (Fig.6.4Ciii) suggest that those origins common to all conditions, which were previously thought to be unaffected by Y RNA depletion, were in fact less active when Y RNAs were removed. The removal of Y RNAs resulted in a reduced probability of origin firing/initiation, implying that Y RNAs are involved in stimulating DNA replication origin firing. Consequently, those low relative activity ds-iniSeq origins present in the mock-depleted & xNuRD overlap grouping (Fig.6.4Cii) were sufficiently low that the further activity reduction upon Y RNA depletion resulted in them not firing sufficiently frequently to be called as origins; resulting in their disappearance when Y RNAs were removed.

I used the list of 10,357 mock-depleted origins as a mask for the Y RNA-depleted and xNuRD addition conditions, to assess the impact of Y RNA depletion and xNuRD addition on origins specifically found to be active in the mock-depleted condition. This assessment encompassed the relative origin activities ( $HL/(HL+LL)$ ) of these mock-depleted origin sites in all experimental conditions for direct comparison (Fig.6.5).



**Figure 6.5:** The mock-depleted ds-*iniSeq* origins were applied as a “mask” to the other two experimental conditions. The mock-depleted mask was used in SeqMonk to define the mock-depleted origin sites for determination of relative site activities in the Y RNA-depleted and xNuRD addition experimental condition. The normalised read counts of the HL and LL DNA, in the Y RNA-depleted and xNuRD addition conditions at these sites were quantified at the mock-depleted mask sites. Their relative activities were calculated as normal (HL/ (HL+LL)). This allowed for the assessment of the impact of Y RNA depletion and xNuRD addition on those ds-*iniSeq* origins present in the mock-depleted experimental condition. (A) Shows the origin activities of the mock-depleted ds-*iniSeq* origins and the origin activities of these same mock-depleted mask sites present in the Y RNA-depleted and xNuRD addition conditions. The activities of the random sites corresponding to the mock-depleted origins are also shown (ochre). (B) The scatter of the ds-*iniSeq* origin activities that overlapped one another in the following combinations: (i) mock-depleted (x axis) vs mock-depleted mask present in Y RNA-depleted (y axis); (ii) mock-depleted mask present in xNuRD addition (x axis) vs mock-depleted mask present in Y RNA-depleted (y axis); and (iii) mock-depleted (x axis) vs mock-depleted mask present in xNuRD addition (y axis). Linear regressions, Pearson’s test for correlation and ANOVA tests for significance were conducted. The correlational R value and the ANOVA test result are indicated on each plot. The dark red dashed line showed the perfect correlation ( $R = 1$ ) for comparative purposes.

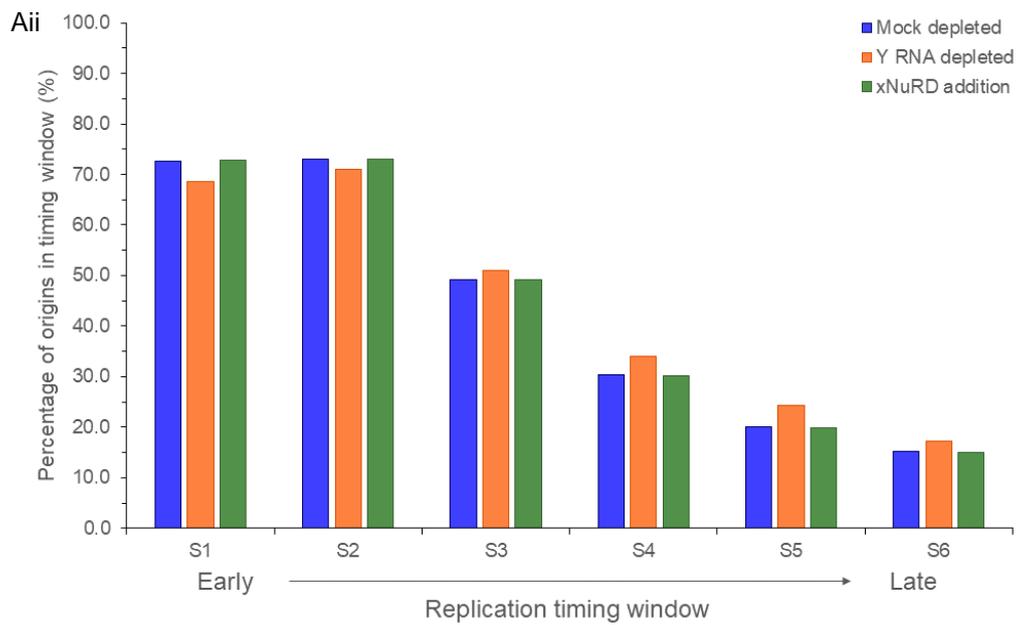
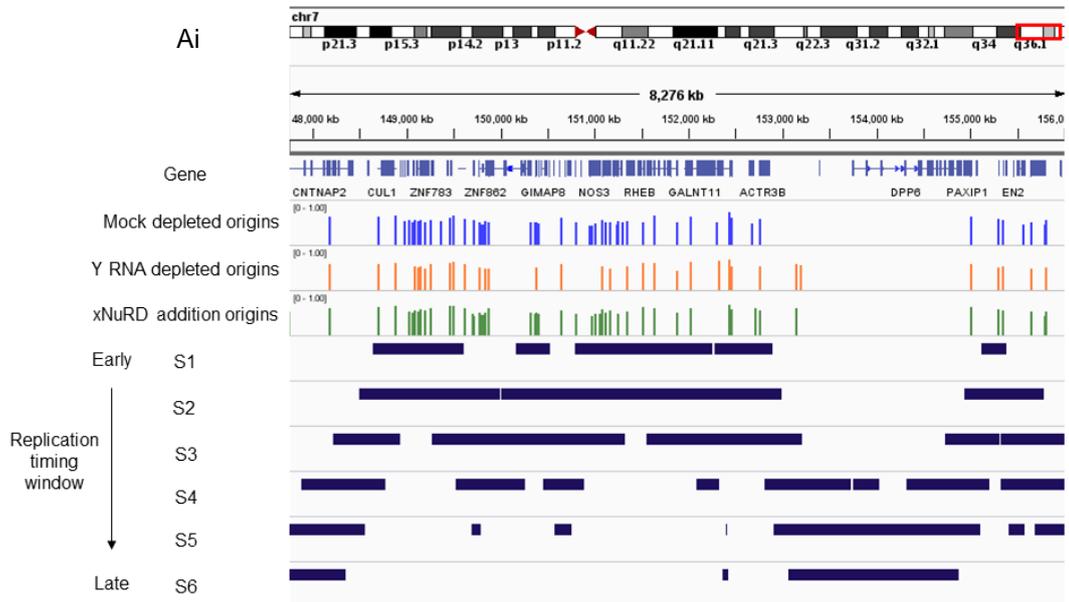
The relative activities of the mock-depleted origins in all three conditions were significantly greater than that of the corresponding randomised sites (Fig.6.5A). The mock-depleted origins showed the statistically highest origin activities. The same sites in the xNuRD addition condition showed similar but statistically lower activities ( $p=1.34*10^{-5}$ ) with a larger range. By contrast, the mock-depleted origin sites in the Y RNA-depleted condition possessed the significantly lowest activities ( $p < 2.2*10^{-16}$ ) when compared to the same sites in the mock-depleted and xNuRD addition conditions. Additionally, the range of these mock-depleted origin sites in the Y RNA-depleted condition was marginally larger and much larger than the xNuRD addition and mock-depleted conditions respectively. This observation of the reduction in relative activity of those origins identified in the mock-depleted condition has added further weight to the argument that Y RNAs are involved in the firing/initiation activity of replication origins.

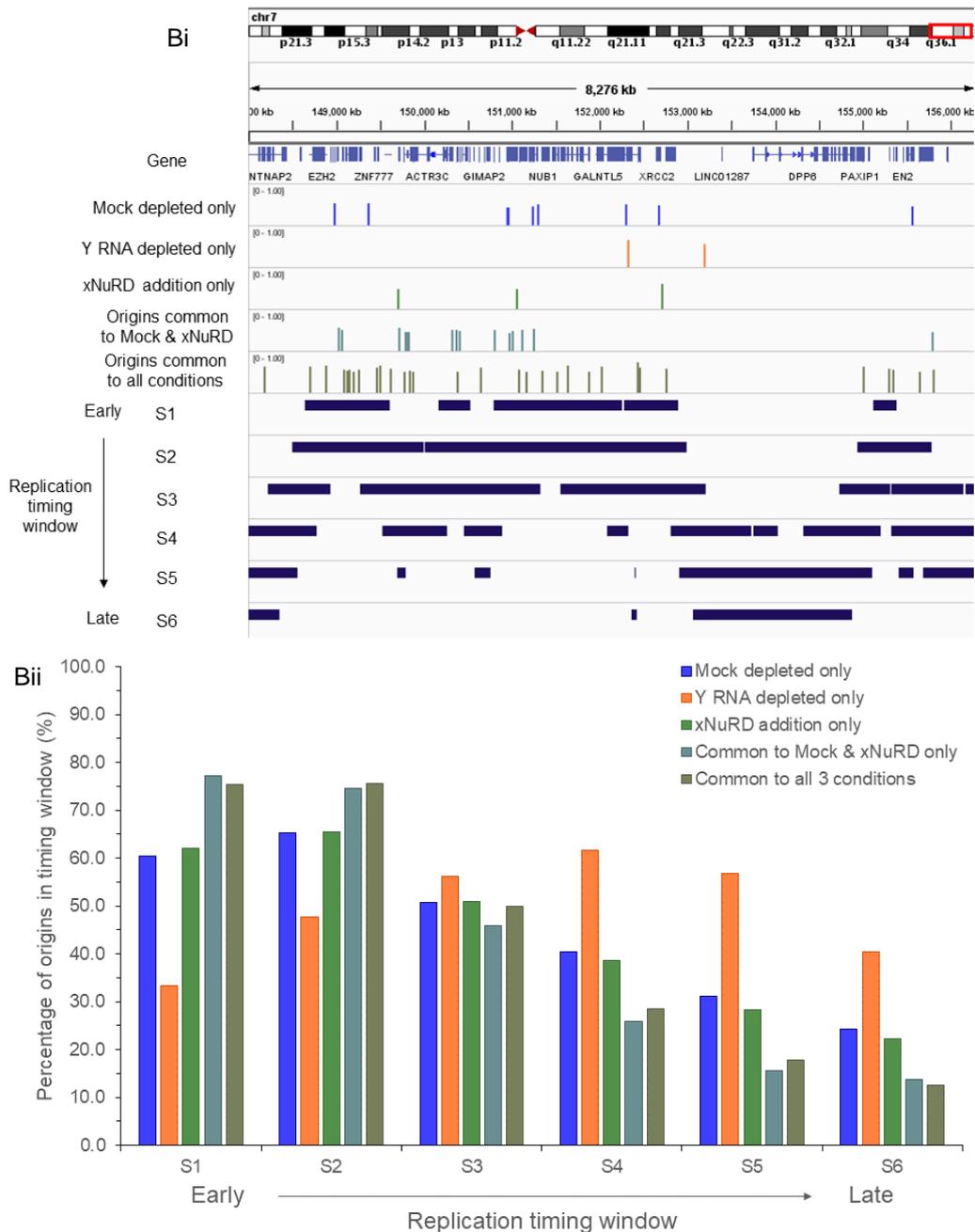
The degree of correlation of activities of the mock-depleted origin sites were compared (Fig.6.5B) in the: mock-depleted vs Y RNA-depleted (Fig.6.5Bi); mock-depleted vs xNuRD addition (Fig.6.5Bii); and the mock-depleted vs xNuRD addition (Fig.6.5Biii) were all strongly positive ( $R$  0.84, 0.83 & 0.82). The activities of the Y RNA-depleted origins were reduced when compared to the other conditions. Again, this data suggests that Y RNA depletion reduced origin activity of all origins uniformly.

#### *6.2.4 How do Y RNAs and xNuRD impact DNA replication origins? – replication timing*

The control origins showed a distinct preference for replication timing windows associated with early replication (Fig.5.6). To determine whether Y RNA depletion and xNuRD addition impacted the timing windows the ds-iniSeq origins were present in, I investigated the distribution of the 10,357 mock-depleted, 7,076 Y RNA-depleted and 10,329 xNuRD addition origins throughout DNA replication (Fig.6.6A), by comparison with the repli-seq data from HeLa cells (8).

To investigate if the timing of particular subgroups of origins are affected specifically, I evaluated the distribution of ds-iniSeq origins in each replication timing window from the following overlap groupings (determined in Fig.6.4): mock-depleted only (Y RNA dependent & not rescued by xNuRD), Y RNA-depleted only (Y RNA independent), xNuRD addition (unique to xNuRD), Mock & xNuRD (removed by Y RNA depletion & rescued by xNuRD), and common to all conditions (reduced origin activities with Y RNA depletion which were rescued by xNuRD) (Fig.6.6B).





**Figure 6.6:** The comparison of ds-inciSeq origins of various conditions with replication timing profiles/windows S1 (early replicating) to S6 (late replicating) from HeLa cells (8). (Ai) The mock-depleted (blue), Y RNA-depleted (orange) and xNuRD addition (green) ds-inciSeq origins, and the replication timing profiles (indigo) at a region on chromosome 7, were visualised on the IGV. (Aii) The percentage of the mock-depleted (blue), Y RNA-depleted (orange) and xNuRD addition (green) ds-inciSeq origins, which were found in each replication timing window. (Bi) The ds-inciSeq origins present in the mock-depleted only (blue), Y RNA-depleted only (orange), xNuRD addition only (green), common to mock-depleted and xNuRD addition (teal) and common to mock-depleted, Y RNA-depleted and xNuRD addition (all 3 conditions) (olive) overlap groupings and the replication timing profiles (indigo) at a region on chromosome 7, were visualised on the IGV. (Bii) The percentage of ds-inciSeq origins present in the mock-depleted only (blue), Y RNA-depleted only (orange), xNuRD addition only (green), common to mock-depleted and xNuRD addition (teal), and common to mock-depleted, Y RNA-depleted and xNuRD addition (all 3 conditions) (olive) overlap groupings, which were found in each replication timing window.

The representative IGV image of all the mock-depleted, Y RNA-depleted and xNuRD addition origins and the corresponding replication timing windows showed a negligible difference on the distribution of these origins in each window (Fig.6.6Ai). The vast majority of the origins from all three conditions were predominantly present in earlier replication timing windows (S1-3).

The genome-wide assessment of the distribution of mock-depleted, Y RNA-depleted and xNuRD addition origins in each replication timing window (Fig.6.6Aii) agreed with the observations in Fig.6.6Ai and were very similar to the control origins (Fig.5.6). The vast majority of the mock-depleted, Y RNA-depleted and xNuRD addition origins were present in early replication timing windows (S1 68.5-72.9% & S2 70.9-73.1%) with very few in late replication timing windows (S5 19.8-24.2%; S6 15.1-17.3%). The percentage of these ds-iniSeq origins present in the mid-replication timing windows decreased rapidly (S3 49.2-51.0%; S4 30.2-34.0%). There was a negligible difference in the distribution of origins in each replication timing window for all the mock-depleted, Y RNA-depleted and xNuRD addition origins. All the Y RNA-depleted origins showed a very slight but consistent shift towards later replication timing windows.

For a more detailed assessment of Y RNA removal and xNuRD addition on the distribution of ds-iniSeq origins in replication timing windows, I compared the selected overlap groupings with the timing windows. The representative IGV imaged showed ds-iniSeq origins from each overlap grouping with the separate replication timing windows (Fig.6.6Bi). The larger number of origins common to all conditions and those common to mock-depleted & xNuRD addition were most frequently found in earlier timing windows. The fewer mock-depleted only and xNuRD addition only origins were predominantly in earlier timing windows. However, the very few Y RNA-depleted only origins were found in both early and later timing windows.

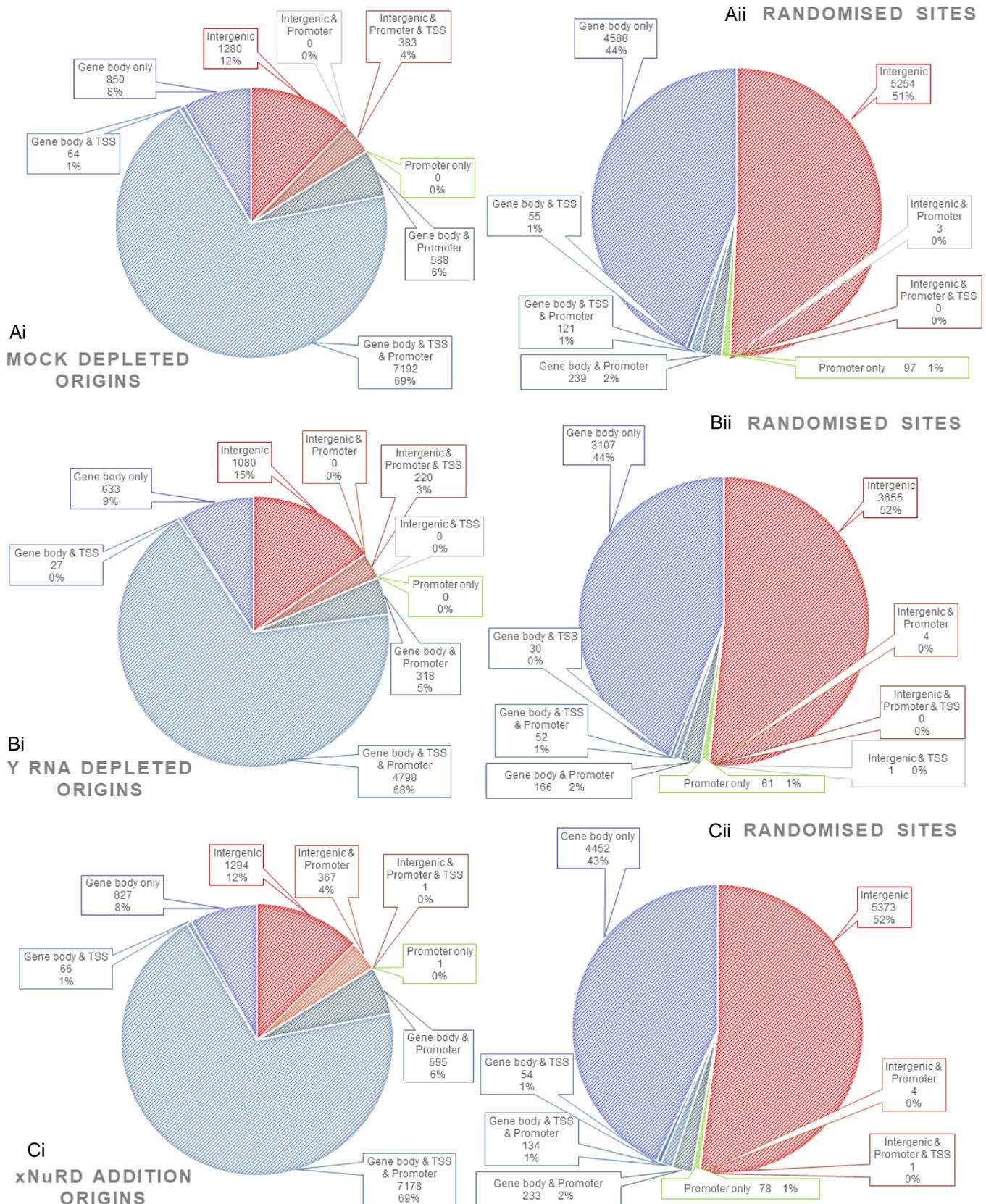
I determined the replication timing window of these ds-iniSeq origins in each overlap grouping on a genome-wide scale (Fig.6.6Bii) and found that the origins common to all conditions and common to mock-depleted & xNuRD addition conditions most closely resembled that of the control origins (Fig.5.6B). The origins were predominantly found in early replicating domains (S1=75.4% & 77.3%; S2=75.5% & 74.6%), with a rapid reduction in origins present across mid S-phase (S3=49.9% & 45.9%; S4=28.5% & 25.9%) and a small occupancy at late timing windows (S5=17.8% & 15.6%; S6=12.6%&13.7%).

The ds-iniSeq origins in each replication timing window of the mock-depleted only and xNuRD addition only overlap groupings followed a similar distribution with a slight enrichment in later replicating domains to that of the common to all and mock-depleted & xNuRD addition origins.

The Y RNA-depleted only origins showed the greatest difference in the occupancy of each replication timing window. Most of these origins were present in the mid-late replication timing windows (S3=56.2%; S4=61.7%; S5=56.7%; S6=40.4%). There were far fewer Y RNA-depleted only origins present in the early timing windows (S1=33.3% & S2=47.6%) compared to the ds-iniSeq origins from the other overlap groupings, demonstrating that the 760 highly active Y RNA-depleted only origins were present at regions of DNA associated with late DNA replication.

#### *6.2.5 How do Y RNAs and xNuRD impact DNA replication origins? – Genomic and epigenetic features*

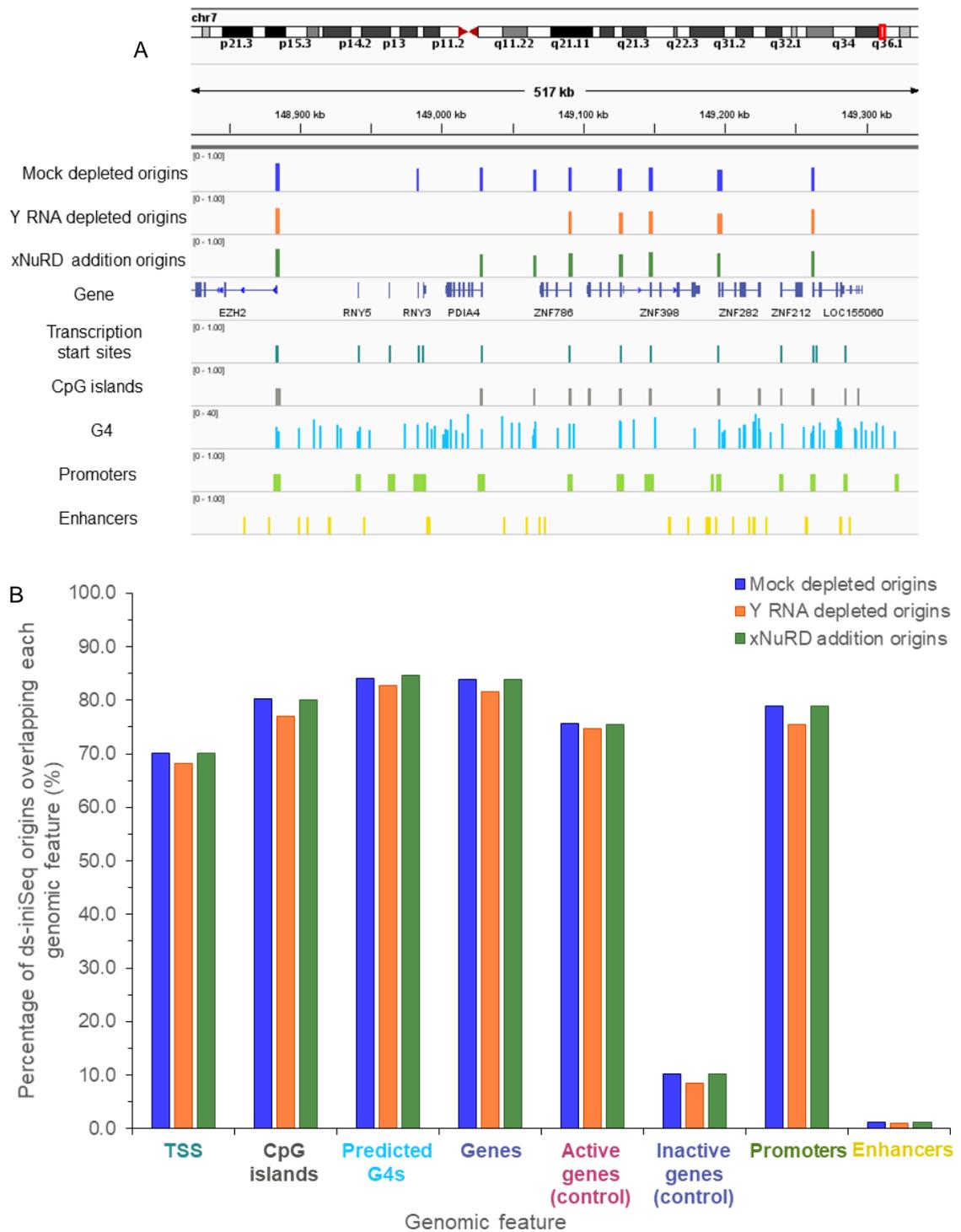
As demonstrated in chapter 5 and previous literature, DNA replication origins are associated with genome features including TSS, CGIs, G4s, genes and promoters. I performed overlap analyses to assess whether Y RNA removal and xNuRD addition influenced the association of replication origins with these genomic features. I assessed the occupancy of all the mock-depleted, Y RNA-depleted and xNuRD addition origins and corresponding random sites (for comparison) at gene bodies (with/without TSS and/or promoters) and intergenic DNA (Fig.6.7).



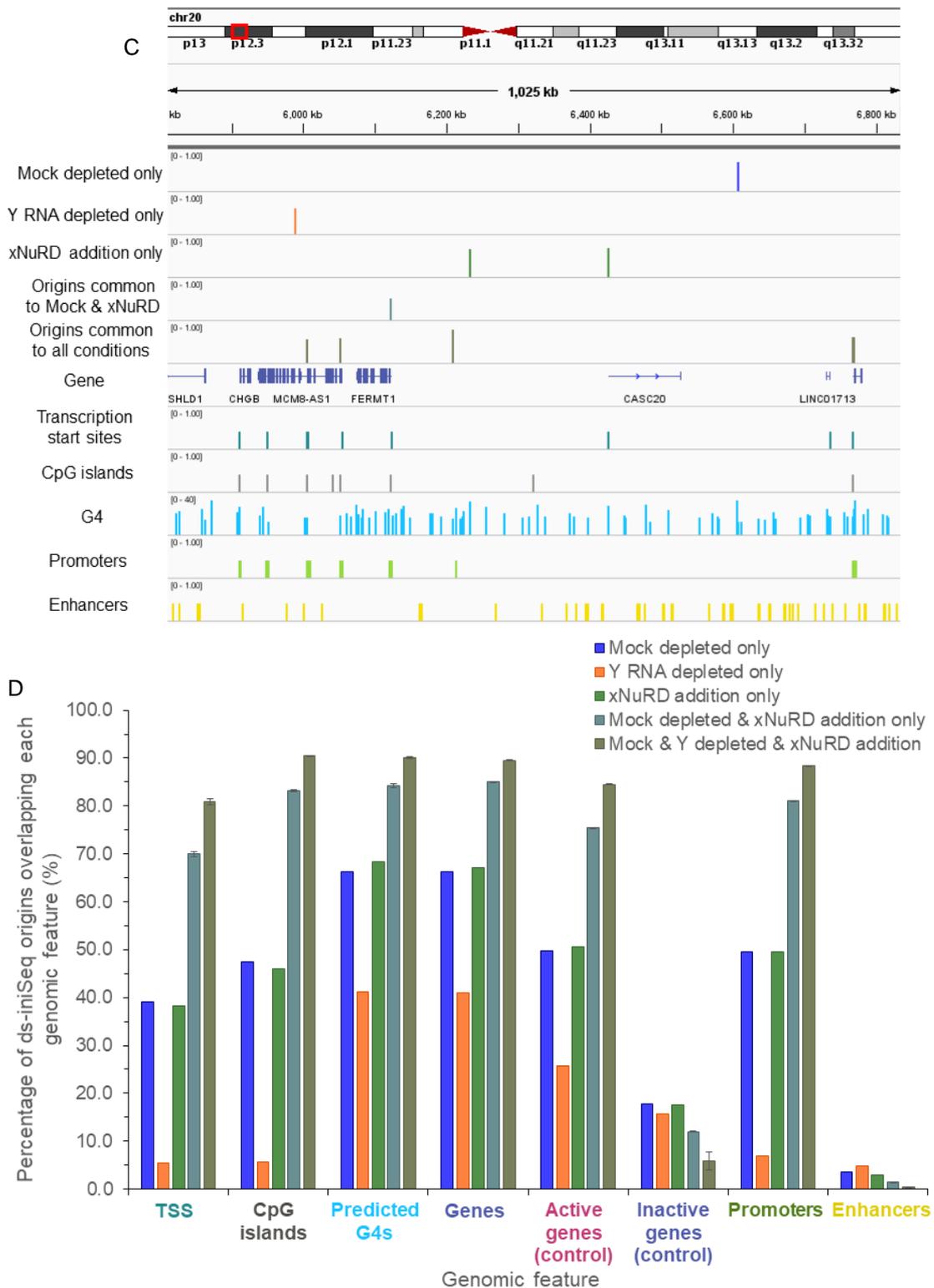
**Figure 6.7:** (A) Pie chart showing the numbers & percentages of mock-depleted ds-iniSeq origins (i) and the corresponding random sites (ii) that overlap with gene bodies (blue colours) (including genes, transcription start sites and/or promoters) and intergenic DNA regions (red colours) (including intergenic DNA, transcription start sites and/or promoters). (B) Pie chart showing the numbers & percentages of Y RNA-depleted ds-iniSeq origins (i) and the corresponding random sites (ii) that overlap with gene bodies (blue colours) (including genes, transcription start sites and/or promoters) and intergenic DNA regions (red colours) (including intergenic DNA, transcription start sites and/or promoters). (C) Pie chart showing the numbers & percentages of xNuRD addition ds-iniSeq origins (i) and the corresponding random sites (ii) that overlap with gene bodies (blue colours) (including genes, transcription start sites and/or promoters) and intergenic DNA regions (red colours) (including intergenic DNA, transcription start sites and/or promoters).

I found that 84% of all mock-depleted (Fig.6.7Ai), 82% of all Y RNA-depleted (Fig.6.7Bi) and 84% of all xNuRD addition (Fig.6.7Ci) origins associated with gene bodies (and/or TSS and/or promoters). Of the corresponding random sites for the mock-depleted (Fig.6.7Aii), Y RNA-depleted (Fig.6.7Bii) and xNuRD addition (Fig.6.7Cii) conditions, 48%, 47% and 48%, respectively, overlapped with gene bodies (and/or TSS and/or promoters). This indicates that origins from all conditions were specifically enriched at gene bodies (and/or TSS and/or promoters), and depleted at intergenic DNA. The overlap of the mock-depleted, Y RNA-depleted and xNuRD addition origins with all genetic elements were almost identical to that of the control origins (Fig.5.8), as were the corresponding random sites.

I compared the mock-depleted, Y RNA-depleted and xNuRD addition ds-iniSeq origins and the ds-iniSeq origins from my selected overlap groupings with the following genomic features: TSS, CGIs, predicted G4s, genes (active and inactive), promoters and enhancers (Fig.6.8).



**Figure 6.8:** (A) An IGV image of the mock-depleted (blue), Y RNA-depleted (orange) and xNuRD addition (green) ds-iniSeq and the genomic features, genes, transcription start sites, CpG islands, G4s, promoters and enhancers, at the region near the EZH2 gene. (B) The percentage of the mock-depleted (blue), Y RNA-depleted (orange) and xNuRD addition (green) ds-iniSeq origins which were present at/ overlapped with the following genomic features: transcription start sites (TSS), CpG islands; predicted G4s; genes; active & inactive genes (determined by SeqMock RNA-seq quantitation pipeline conducted on untreated EJ30s cells; active genes possessed a Reads per kilobase of transcript per million reads (RPKM) of < 0 and inactive genes possessed a RPKM of 0), promoters and enhancers.



**Figure 6.8:** (C) An IGV image of ds-iniSeq origins present in the mock depleted only (blue), Y RNA-depleted only (orange), xNuRD addition only (green) common to mock depleted and xNuRD addition (teal), and common to mock depleted, Y RNA-depleted and xNuRD addition (all 3 conditions) (olive) overlap groupings and the genomic features, genes (blue-purple), transcription start sites, CpG islands, G4s, promoters and enhancers, at the region near the EZH2 gene. (D) The percentage of ds-iniSeq origins present in the mock depleted only (blue), Y RNA-depleted only (orange), xNuRD addition only (green), common to mock depleted and xNuRD addition (teal), and common to mock depleted, Y RNA-depleted and xNuRD addition (all 3 conditions) (olive) overlap groupings, which were present at/ overlapped with the following genomic features: transcription start sites (TSS); CpG islands; predicted G4s; genes; active & inactive genes (determined by SeqMock RNA-seq quantitation pipeline conducted on untreated EJ30s cells; active genes possessed a RPKM of < 0 and inactive genes possessed a RPKM of 0), promoters and enhancers.

The representative IGV image showed the positions of all the mock-depleted, Y RNA-depleted and xNuRD addition ds-iniSeq origins when compared to the above genomic features (Fig.6.8A). All origins for each condition were associated with at least one of the genomic features (TSS, CGIs, G4s, genes & promoters) that were previously found to colocalise with the control origins with high origin activities. As with the control origins, these origins appeared to not overlap with enhancers.

On the genome-wide scale, origins for the mock-depleted, Y RNA-depleted and xNuRD addition conditions highly colocalised with TSS (70.1%, 68.2%, 70.1% respectively), CGIs (80.2%, 77.0%, 80.0% respectively), predicted G4s (84.1%, 82.7%, 84.6% respectively), genes (83.9%, 81.6%, 83.9% respectively), in particular active genes (75.6%, 74.8%, 75.5% respectively), and promoters (78.8%, 75.4%, 78.8% respectively) (Fig.6.8B). All the origins from each condition were poorly associated with inactive genes (10.1%, 8.5%, 10.2% respectively) and enhancers (1.3%, 1.0%, 1.1% respectively). These findings were consistent with the levels of colocalisation with control origins with these genomic features (Fig.5.9&11&12). Although the mock-depleted, Y RNA-depleted and xNuRD addition origins showed very similar levels of overlap with these genomic features, in all cases slightly fewer Y RNA-depleted origins colocalised with these features compared to the other experimental conditions. This indicated that overall, the removal of Y RNAs and the addition of xNuRD had, at most, a marginal effect on the association ds-iniSeq origins with genomic features.

I examined overlap levels of ds-iniSeq origins from my specified overlap groupings with these genomic features. The IGV image shows a representative genomic domain with these overlap grouping origins with the genomic features (Fig.6.8C). Those ds-iniSeq origins present in mock-depleted & xNuRD addition conditions and common to all conditions were associated with at least one of the following genomic features; TSS, CGIs, predicted G4s, genes or promoters. The single mock-depleted only origin overlapped with a predicted G4. Of the two xNuRD addition only origins shown here, one overlapped with a G4 and a promoter, and the other overlapped with a TSS and G4. By contrast, the single Y RNA-depleted only origin was not associated with any of these genomic features.

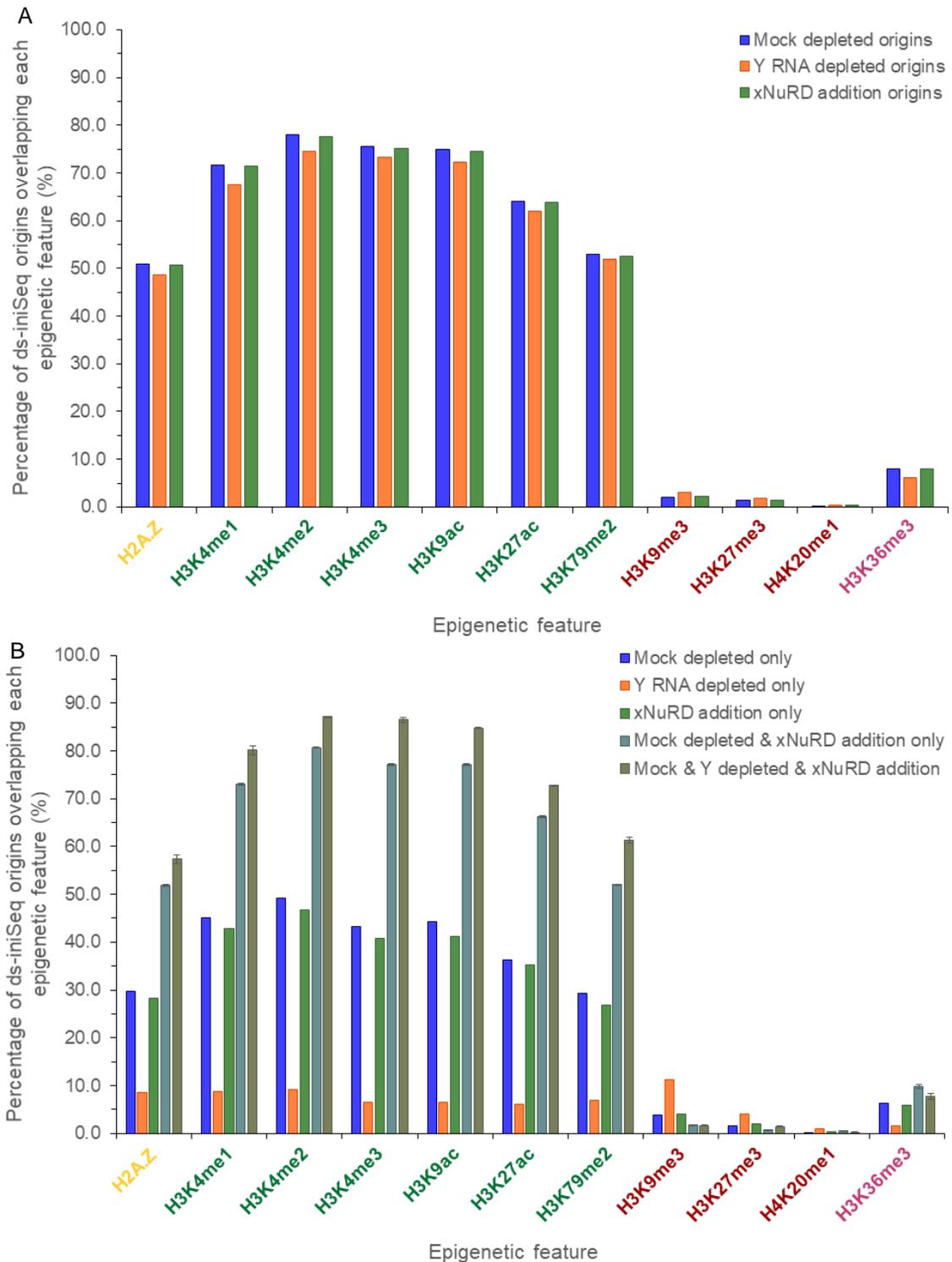
On the genome-wide scale, the ds-iniSeq origins of the common to all conditions overlap grouping possessed the highest degree of overlap with TSS (80.9%), CGIs (90.5%), predicted G4s (90.1%), genes (89.6%) of which 84.6% overlapped with active genes and promoters (88.4%) (Fig.6.8D). They also had the lowest association with inactive genes (5.9%) and enhancers (0.4%). Those origins present in the mock-depleted & xNuRD addition overlap grouping possessed a similarly high association with TSS (70.0%), CGIs (83.1%), predicted G4s (84.3%), genes (85.1%), active genes (75.4%), and promoters (81.0%) that was slightly lower than the common to all origins. The occupancy of mock-depleted & xNuRD

addition origins at inactive genes (12.0%) and enhancers (1.5%) was slightly greater than the common to all origins.

Ds-*ini*Seq origins present in the mock-depleted only and xNuRD addition only overlap grouping possessed very similar levels of colocalisation with all the genomic features, to one another. Occupancy of these origins at TSS (39.0%, 38.2% respectively), CGIs (47.5%, 46.1% respectively), predicted G4s (66.2%, 68.4% respectively), genes (66.3%, 67.2% respectively), active genes (49.8%, 50.7% respectively), and promoters (49.5%, 49.5% respectively) was almost half that of the common to all conditions and mock-depleted & xNuRD addition origins. The mock-depleted only and xNuRD addition only origins showed an increased occupancy at inactive genes (17.7%, 17.5%) and enhancers (3.6%, 2.9%), compared that of the common to all conditions and mock-depleted & xNuRD addition origins.

The Y RNA-depleted only origins showed almost the inverse relationship of occupancy at the genomic features compared to the origins of the other overlap groupings. The colocalisation of the Y RNA-depleted only origins at TSS (5.5%), CGIs (5.7%) and promoters (6.8%) was incredibly low and colocalisation at G4s (41.2%), genes (40.9%)/active genes (25.7%) were substantially reduced, compared to that of the other overlap condition origins. The occupancy of the Y RNA-depleted only origins at inactive genes was marginally reduced, compared to mock-depleted only and xNuRD addition only origins, and marginally increased compared to the mock-depleted & xNuRD addition, and common to all conditions origins. The association of the Y RNA-depleted only origins at enhancers (4.9%) was increased compared to all other overlap groupings. Interestingly these Y RNA-depleted only origins possessed high origin activities (Fig.6.4Ci) but were not strongly colocalised with any genomic feature associated with high origin activity in the control condition (Fig.5.9&11&12).

I compared the mock-depleted, Y RNA-depleted and xNuRD addition origins and my specified overlap groupings with the epigenetic features I investigated in chapter 5 (Fig.6.9). I used the ChIP-seq data generated, in the HCT116 cell line for; the H2A.Z histone isoform (associated with replication origins), the histone marks associated with early firing replication origins (H3K4me1/2/3, H3K9ac, H3K27ac and H3K79me2), the histone marks associated with late firing replication origins (H3K9me3, H3K27me3 and H4K20me1), and a histone mark for transcriptional activity (H3K36me3).



**Figure 6.9:** (A) The percentage of the mock depleted (blue), Y RNA-depleted (orange) and xNuRD addition (green) ds-iniSeq origins, which were present at/ overlapped with the following epigenetic features (obtained from ChIP-seq data conducted on HCT116 cells (Encode): H2A.Z (associated with replication origins), H3K4me1/2/3, H3K9ac, H3K27ac, H3K79me2 (associated with early firing replication origins), H3K9me3, H3K27me3, H4K20me1 (associated with late firing replication origins) and H3K36me3 (associated with active transcription). (B) The percentage of ds-iniSeq origins present in the mock depleted only (blue), Y RNA-depleted only (orange), xNuRD addition only (green), common to mock depleted and xNuRD addition (teal), and common to mock depleted, Y RNA-depleted and xNuRD addition (all 3 conditions) (olive) overlap groupings, which were present at/overlapped with the following epigenetic features (obtained from ChIP-seq data conducted on HCT116 cells (Encode): H2A.Z (associated with replication origins), H3K4me1/2/3, H3K9ac, H3K27ac, H3K79me2 (associated with early firing replication origins), H3K9me3, H3K27me3, H4K20me1 (associated with late firing replication origins) and H3K36me3 (associated with active transcription).

The overlap of all mock-depleted, Y RNA-depleted and xNuRD addition ds-*iniSeq* origins with various epigenetic features showed that ~50% of the origins in all three conditions associated with the H2A.Z histone isoform (Fig.6.9A). There was a high degree of overlap between the mock-depleted, Y RNA-depleted and xNuRD addition origins with the early firing origin-associated (EFOA) marks, H3K4me1 (71.7%, 67.6%, 71.4% respectively), H3K4me2 (78.0%, 74.5%, 77.6% respectively), H3K4me3 (75.6%, 73.4%, 75.1% respectively), H3K9ac (74.9%, 72.2%, 74.4% respectively), H3K27ac (64.0%, 61.9%, 63.9% respectively) and H3K79me2 (52.9%, 51.9%, 52.6% respectively). Despite this, there were slightly fewer Y RNA-depleted origins that overlapped with them compared to the origins of other two conditions.

By contrast, the mock-depleted, Y RNA-depleted and xNuRD addition origins overlapped poorly with those histone marks associated with late firing origins, namely H3K9me3 (2.1%, 3.0%, 2.2% respectively), H3K27me3 (1.3%, 1.9%, 1.4% respectively) and H4K20me1 (0.3%, 0.4%, 0.3% respectively). The mock-depleted, Y RNA-depleted and xNuRD addition origins were not well associated with the active transcription mark, H3K36me3 (8.0%, 6.1%, 8.0% respectively). Slightly fewer Y RNA-depleted origins overlapped H3K36me3 and slightly more Y RNA-depleted origins overlapped H3K9me3 and H3K27me3, compared to the origins of the other conditions. The levels of overlap between these mock-depleted, Y RNA-depleted and xNuRD addition origins with the epigenetic features were highly similar to the overlap found in the control origins (Fig.5.13). As with the genomic features, these data showed that Y RNA removal and xNuRD addition had, at most, a marginal effect on the association of active ds-*iniSeq* origins with these epigenetic features.

Finally, I compared the association of the origins of my specified overlap groupings with the same epigenetic features on a genome-wide scale (Fig.6.9B).

Ds-*iniSeq* origins common to all conditions overlap grouping possessed the highest overlap with H2A.Z (57.3%) and the EFOA histone marks, H3K4me1 (80.1%), H3K4me2 (87.1%), H3K4me3 (86.5%), H3K9ac (84.8%), H3K27ac (72.8%) and H3K79me2 (61.3%). These origins also possessed the lowest association with the late firing origin associated (LFOA) histone marks, H3K9me3 (1.7%), H4K20me1 (0.2%) and the second lowest association with H3K27me3 (1.4%).

Ds-*iniSeq* origins present in the mock-depleted & xNuRD addition overlap grouping colocalised with H2A.Z (51.9%) and the EFOA histone marks, H3K4me1 (73.0%), H3K4me2 (80.6%), H3K4me3 (77.1%), H3K9ac (77.1%), H3K27ac (66.2%) and H3K79me2 (52.1%) at slightly lower levels than the common to all origins. These origins also possessed similar overlaps to the common to all origins, with the LFOA histone marks, H3K9me3 (1.7%), H3K27me3 (0.8%) and H4K20me1 (0.5%).

These mock-depleted & xNuRD addition and common to all conditions origins possessed high activities and a high level of association with H2A.Z and EFOA histone marks. This observation was consistent with the previous observations in the control ds-iniSeq origin, of the impact of colocalisation with these features on origin activity and subsequently the probability of firing/initiation (Fig.5.13/14).

The ds-iniSeq origins present in the mock-depleted only and xNuRD addition only overlap grouping possessed very similar levels of overlap with all the epigenetic features, to one another. These origins overlapped substantially with H2A.Z (29.8%, 28.4%), H3K4me1 (45.1%, 42.9%), H3K4me2 (49.2%, 46.8%), H3K4me3 (43.3%, 40.7%), H3K9ac (44.2%, 41.2%), H3K27ac (36.3%, 35.2%) and H3K79me2 (29.3%, 26.7%), but overlapped poorly with H3K9me3 (3.8%, 4.0%), H3K27me3 (1.6%, 2.0%) and H4K20me1 (0.1%, 0.2%). Overall the occupancy of these origins with H2A.Z and the EFOA histone marks was approximately half that of the common to all conditions origins.

The Y RNA-depleted only origins showed a very low association with all of these epigenetic features. The overlap of these origins with H2A.Z (8.6%), H3K4me1 (8.8%), H3K4me2 (9.2%), H3K4me3 (6.4%), H3K9ac (6.6%), H3K27ac (6.1%) and H3K79me2 (7.0%) was substantially reduced compared to the origins present in the other overlap groupings. However, the overlap of the Y RNA-depleted only origins with H3K9me3 (11.3%) was considerably increased, and the overlap with H3K27me3 (4.1%) and H4K20me1 (0.9%) was moderately increased, compared to the origins present in the other overlap groupings. These Y RNA-depleted only origins possessed high origin activities but were not strongly colocalised with H2A.Z and the EFOA histone marks that were previously linked to high origin activity in the control origins. This was similar to the findings of these Y RNA-depleted only origins relationship with those genomic features previously associated with high origin activity (Fig.6.8). These Y RNA-depleted only origins may represent late firing origins that become activated earlier during the *in vitro* replication reaction, in the absence of Y RNAs. Alternatively, they may represent a group of normally dormant origin sites that become active upon the removal of Y RNAs.

As expected from the analysis of the control origins with the active transcription mark H3K36me3 (Fig.5.13), the origins of the common to all conditions (7.7%), mock-depleted & xNuRD addition (9.7%), mock-depleted only (6.4%) and xNuRD addition only (5.8%) overlap groupings colocalised poorly with H3K36me3. Interestingly, the overlap of the Y RNA-depleted only (1.6%) origins with H3K36me3 was reduced even further compared to the origins present in the other overlap groupings.

### 6.3 Discussion

In this chapter, I addressed the following questions: How do Y RNAs affect human replication origin specification and activation? And how does xNuRD affect human replication origin specification and activation in the absence of Y RNAs?

Y RNAs have been identified as essential DNA replication initiation factors both *in vitro* and *in vivo*. Upon their removal in the human cell-free system, DNA replication was drastically and significantly reduced (Fig.6.1) and Y RNA removal *in vivo* resulted in the inhibition of DNA replication and cell proliferation (1,4,9,10). As Y RNAs are essential in replication initiation but not elongation (11), it is possible that they may be involved in replication origin specification and/or activation.

The data shown here demonstrated that Y RNAs impacted the activation of replication origins. Upon Y RNA removal, fewer origins fired and those that did fire had lower relative origin activities. The mock-depleted origins that no longer fired in the absence of Y RNAs possessed lower activities than those that did (Fig.6.4). These suggest a clear role for Y RNAs in the activation of replication origins. As relative origin activity is a proxy for probability of an origin firing, I surmise that Y RNAs increase the probability of a given origin firing.

The subunit isoform composition of the 6-subunit chromatin remodelling complex NuRD determines its function and cell type specificity (12–14). A specific xNuRD comprised of CHD4, MTA2, RBBP7/RbAp46, MBD3, p66 $\alpha$  and HDACm (and/or HDAC1), has been identified as a DNA replication initiation factor that can functionally substitute for hY RNAs. This subunit composition is specifically found in *X. laevis* embryos prior to the mid-blastula transition (MBT) and is presumably responsible for its function in DNA replication (2). During pre-MBT, *X. laevis* embryos undergo rapid DNA replication cycles and no sequence specificity for initiation sites (3,6). This xNuRD complex functionally substitutes for Y RNA in human cell nuclei (2), which allowed me to answer the question “does xNuRD cause delocalised DNA replication initiation?”.

The effect of xNuRD on human replication origin firing through assessment by ds-iniSeq analysis showed that the presence of xNuRD did not cause delocalised DNA replication but resulted in the firing of discrete replication origin sites (Fig.6.3). Additionally, xNuRD mostly re-established the firing and relative origin activities of those origins affected by Y RNA removal, suggesting that xNuRD targets the same discrete sites as hY RNAs (Fig.6.4).

Traditionally NuRD is associated with transcriptional silencing (13,15). In chapter 5, I determined that the absence of transcriptional/gene activity was associated with low relative origin activity (Fig.5.11D). The observation that the presence of xNuRD increased relative origin activity was unexpected due to NuRD's previous association with transcriptional

silencing. However, there are 2 mutually exclusive MBD isoforms present in different NuRD complexes (16); MBD2 binds with methylated CpG sequences and promotes transcriptional silencing, whereas MBD3 binds unmethylated CGIs and is enriched at active promoters (17). Both CGIs/CGI-promoters and active genes (thus implying active promoters) were associated with high relative origin activity (Fig.5.9G/H&12C) and xNuRD possesses the MBD3 subunit isoform. Furthermore, the inhibition of MBD3 in the human cell-free system, inhibited DNA replication (2). These findings may explain the increase in relative origin activity by xNuRD, despite most research showing an association of NuRD complexes with transcriptional silencing. It may also suggest that xNuRD is directed to CGIs and/or CGI-promoters via MBD3, in order to fulfil its potential role in DNA replication. It is important to note that during pre-MBT in *X. laevis* embryos, there is no active transcription and the association of MBD3 with active promoters was likely to be found post-MBT (3,6,17). Nevertheless, the human cell nuclei that these ds-iniSeq origins were identified from undergo active transcription. Consequently, the enrichment of MBD3 at active promoters may still be relevant. Additionally, xNuRD is required for DNA replication and cell proliferation *in vivo* (2). MBD3's preference for CGIs and potentially promoters may constitute a binding site for xNuRD irrespective of transcriptional activity in *X. laevis*. Ultimately, more research must be conducted to further elucidate the role of xNuRD in *X. laevis* embryos and ascertain whether it has the same effect on origin specification and activity in *X. laevis* as it does in human cells.

These data have also shown that all the mock-depleted origins were affected by Y RNA removal; their relative origin activities reduced, in some cases, so low that they did not fire (Fig.6.4). The origins common to all experimental conditions and those active in mock-depleted and xNuRD addition conditions (ie origins affected by Y RNA removal and rescued by xNuRD), showed similar distributions across replication timing windows (bias firing in early replication) and similar colocalisations with the genomic and epigenetic features examined here (Fig.6.6-9). Those origins common to mock-depleted and xNuRD addition conditions had a slightly lower overlap with genomic features associated with high relative activity and EFOA epigenetic features, and a slightly higher colocalisation with genomic features associated with lower relative activities than those origins common to all conditions (Fig.6.8/9). These colocalisations/associations were consistent with my findings for the control origins shown in chapter 5 (Fig.5.9/11-14).

In addition to the above origins, there were mock-depleted origins that were not rescued by xNuRD (mock-depleted only). These origins had similar relative origin activities to those that were rescued by xNuRD (Fig.6.4). This suggests a difference between origins that were and were not rescued by xNuRD, which is separate from mere origin activity prior to Y RNA removal. The only difference observed in these analyses was that the mock-depleted only

origins were less colocalised with genomic features associated with high relative activity and EFOA epigenetic features, than the mock-depleted and xNuRD addition origins (Fig.6.8/9), implying that xNuRD may target origins colocalised with genomic features associated with high relative origin activities and EFOA epigenetic features. The associations with CGIs, TSS and promoters/active genes were more affected than the other genomic features in the mock-depleted only origins (Fig.6.8/9). In addition to MBD3 binding affinity with CGIs and enrichment at active promoters (17), this may offer a potential avenue for a mechanism by which xNuRD brings about DNA replication origin firing; xNuRD may bind to CGIs (and/or CGI-promoters) to increase the probability of a given origin firing, potentially by altering the chromatin environment to one more conducive with origin firing. There were origins present in only the xNuRD addition condition (xNuRD only) that possessed similar activities and colocalisation with genomic and epigenetic features as the mock only origins (Fig.6.4/8/9). One potential additional explanation for the existence of xNuRD only origins is that there may be another factor at play that facilitates these origins to fire, which is feasible as replication is an essential process for the longevity of an organism's life, so is likely to have redundancy.

Finally, I found a group of origins unique to the Y RNA depletion condition (Y RNA-depleted only) that unexpectedly possessed high relative origin activities (Fig.6.4). These origins were activated in a Y RNA-independent manner and suggest that there may be a set of "fail safe" origins that fire if Y RNAs are unable to fulfil their function in DNA replication. There are many origin sites that become licensed for replication but never undergo origin firing under normal conditions. These dormant origins only fire occasionally to resolve issues that may prevent or impede replication (18–20). I propose that these Y RNA-depleted only origins constitute some of these dormant origins and/or are origins that would normally fire later in replication but are activated earlier in the absence of Y RNAs.

There were major differences in the colocalisation of these Y RNA-depleted only origins with genomic and epigenetic features, compared to origins in all other overlap groupings and the control origins discussed in chapter 5. The Y RNA-independent (Y RNA-depleted only) origins had a low association with G4s, genes and active genes and were not associated with CGIs, promoters and TSS (Fig.6.8B). As CGIs and CGI-promoters have been highlighted as features associated with high relative origin activity (Fig.5.12C) and thus high probability of firing, the lack of association with CGIs and promoters by the Y RNA-independent origins would add further weight to the proposal that they are activated dormant origins. I therefore hypothesise that as Y RNA-dependent origins do and Y RNA-independent origins do not associate with CGIs or promoters, Y RNAs may target or associate with CGIs/CGI-promoter sites where they facilitate the origin firing.

I found that (chapter 5) there appeared to be an interplay between CGIs/promoters with EFOA histone marks (Fig.5.16). Additional data showed almost no association between the

Y RNA-independent origins and the EFOA marks (Fig.6.9B). As with those genomic features normally associated with high origin activity, the lack of association between the Y RNA-independent origins and EFOA marks may indicate that these were dormant fail safe origins which fired when Y RNAs were removed and normal origin firing could not take place. The reduction in the colocalisation of Y RNA-independent origins with EFOA marks was proportionate and no one EFOA mark was more affected than any other. This may support the findings from chapter 5, that the number of EFOA marks had the greatest impact on relative origin activity/probability of firing over any individual EFOA mark (Fig.5.14).

Overall, these data suggest that the previously dormant Y RNA-independent origins were independent of the epigenetic environment associated with origin firing in early replication. Previous research found that MCM2-7 complexes are actively loaded on to the DNA irrespective of chromatin status and origins subsequently fire based on open chromatin structure (21). I hypothesise that the origins that normally fire based on open chromatin and epigenetic environment are those origins dependent on Y RNAs and those rescued by xNuRD following Y RNA depletion, whereas the Y RNA-independent origins fired independently of this type of chromatin environment. Alternatively, the Y RNA-independent origins may be origins that normally fire later in DNA replication but were required to fire earlier as the normally early firing (Y RNA-dependent) origins did not fire, rather than be independent of the epigenetic features normally associated with origin firing. I found that the Y RNA-independent origins were more abundant in later replication windows when compared to all other ds-*ini*Seq origins (Fig.6.6B). Additionally, the Y RNA-independent origins had an increased association with the LFOA mark, H3K9me3, compared to all other ds-*ini*Seq origins in these analyses (Fig.6.9B), which support the suggestion above. The overlap of the Y RNA-independent origins with H3K9me3 was almost double that of other origin groups, but the overall overlap was still very low (<15%) (Fig.6.9B), making this suggestion less likely. However, the possibility remains that the Y RNA-independent origins were a mixture of activated dormant origins and later firing origins that initiated earlier in S-phase.

One question that remains is, what are the mechanisms by which Y RNAs and xNuRD bring about origin firing and subsequent DNA replication? Y RNAs interact with ORC, *cdc6* and *cdt1* and colocalise with ORC, MCM2, *cdt1* and *cdc45* on unreplicated euchromatin during late G1. They do not biochemically interact with the replication fork proteins, MCM2-7 & GINS complexes and DNA polymerase/primase, so are unlikely to be involved in replication fork progression (4,22). They associate with euchromatin throughout the cell cycle but are more abundant during S-phase when compared to G1 or G2/M. Finally, Y RNAs are absent from actively replicating foci (which are comprised of multiple origins and elongating replication forks) but quickly re-associate with euchromatin during mid-late S-phase. This association behaviour is consistent with ORC binding. Overall the literature has suggested

that Y RNAs bring about DNA replication initiation via a common pathway with ORC but are not classical licensing factors as they reassociated with replicated chromatin during mid-late S-phase (23,24).

Furthermore, ORC is required for Y RNA binding to chromatin (4). ORC is also required for MCM2-7 loading to the pre-RC, but MCM2-7 complexes are loaded in excess, irrespective of chromatin status resulting in licensed but dormant origins (18–20,25,26), implying that ORC is also bound to chromatin in excess. Only a fraction of the licensed origins actually fire and that is considered to be dependent on chromatin areas with open structure, conducive with DNA replication/ origin firing (20).

Roberts (27) found that hY RNAs bind to the Protein arginine N-methyltransferase 1 (PRMT1), which is responsible for the asymmetric di-methylation of H3R4 (H3R4me2a); a histone mark associated with transcriptional activity (28–33). PRMT1 was subsequently found to be essential for DNA replication in the human cell-free system, presumably through an interaction with Y RNAs (27). H3R4me2a has recently been found to associate with CGIs, along with H3R2me2 and the EFOA marks H3K4me3 and H3K27ac (34). These data provide a further link between CGIs, Y RNAs and DNA replication.

Therefore, I hypothesise that Y RNAs may increase relative replication origin activities and thus the probability of origins firing, through an interaction with chromatin remodellers, such as PRMT1 and others, that are directed to origin site via an affinity for CGIs and/or CGI-promoters, potentially loaded to chromatin by ORC and subsequently modify histone tails in order to generate a chromatin environment more conducive with DNA replication origin firing. Further work must be carried out to confirm or disprove this hypothesis.

Finally, xNuRD may act in a similar way to Y RNAs to bring about DNA replication initiation. It increased the relative activities of mostly the same origins as Y RNA and is associated with CGIs and/or active promoters (17). NuRD has also been found to interact with and facilitate the recruitment of other chromatin remodellers, including PRC2 (13,35). xNuRD is a relatively new player in DNA replication and a substantial amount of work needs to be conducted to determine the mechanism by which it brings about DNA replication and increases the probability of origins firing. Avenues of research may include: the identification of the histone modifications and chromatin environment that xNuRD generates to facilitate replication; any associations of xNuRD with ORC and/or other replication initiation factors; and the colocalisation of xNuRD binding sites with replication origins.

The mechanisms by which Y RNAs and xNuRD bring about DNA replication initiation and origin firing remain not fully elucidated, but my data presented here has revealed clear hints as to what the mechanisms may be, allowed me to propose a new hypothesis and provided additional avenues for further investigation.

## **Chapter 7: The impact of Y RNAs and xNuRD on replication elongation using density-substitution elongation-site sequencing (ds-eloSeq)**

### **7.1 Introduction**

Previously I described the development of the ds-*iniSeq* method for the identification of human replication origins and the use of ds-*iniSeq* to assess the effect of hY RNAs and xNuRD on replication origin specification and activation.

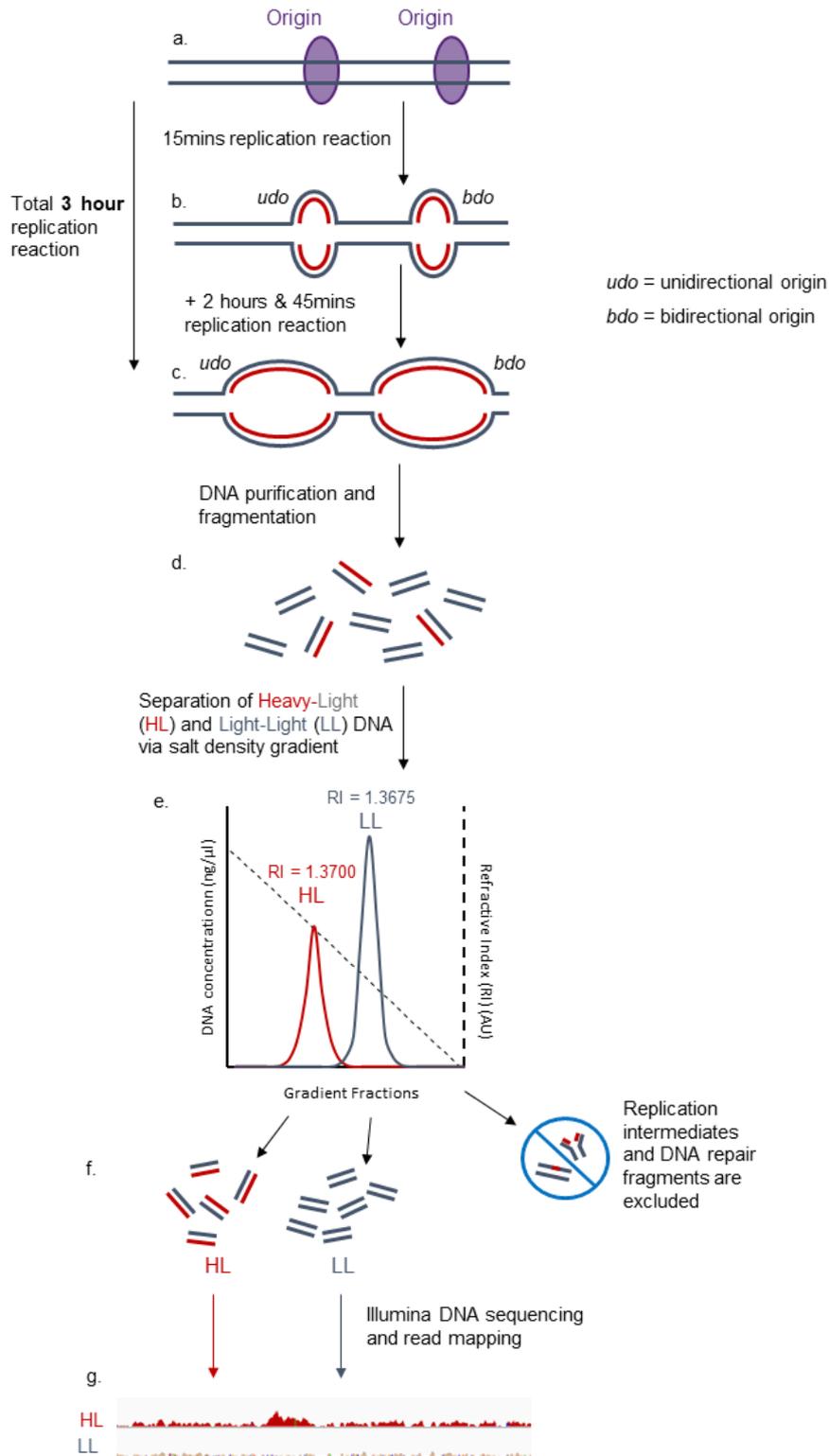
Here, I have adapted ds-*iniSeq* to investigate replication elongation. Density-substitution elongation-site sequencing (ds-*eloSeq*) follows the same method as ds-*iniSeq*, but with an extension of the replication reaction incubation time to 3 hours. I conducted ds-*eloSeq* alongside ds-*iniSeq* to investigate the transition from replication initiation to elongation.

As with the ds-*iniSeq* data presented in chapters 4-6, I performed ds-*eloSeq* reactions, where replication reactions were conducted under control (in the presence of human cytosol containing hY RNAs), mock-depleted (human cytosol treated with a non-human target antisense oligonucleotide), Y RNA-depleted (human cytosol treated with oligonucleotides complementary for hY RNAs, resulting in Y RNA depletion) and xNuRD addition (Y RNA-depleted human cytosol and partially purified xNuRD added) conditions. These allowed me to assess the effect of hY RNAs and xNuRD on DNA replication elongation.

In this chapter, I present a preliminary analysis of the ds-*eloSeq* data conducted in the control, mock-depleted, Y RNA-depleted and xNuRD addition conditions. For the ds-*eloSeq* data analysis, MACS peak caller is inappropriate for calling origins in 3-hour replication incubation time data. MACS was designed for shorter sharper peaks (1), whereas the ds-*eloSeq* data contain wider regions of replicated DNA resulting from replication elongation in addition to the fired origins. I have therefore endeavoured to develop a new method for origin calling from the 3-hour ds-*eloSeq* data, in collaboration with S. Mookerjee (St Catharine's College, Cambridge).

### **7.2 Results and Discussion – Preliminary analysis of ds-eloSeq data**

The ds-*eloSeq* method (schematic in Fig.7.1A) was adapted from the ds-*iniSeq* method, where semi-conservative DNA replication facilitated the incorporation of the heavy BrdUTP into newly synthesised DNA. The replicated Heavy-Light DNA was separated from unreplicated Light-Light DNA via density equilibrium centrifugation and sequenced.



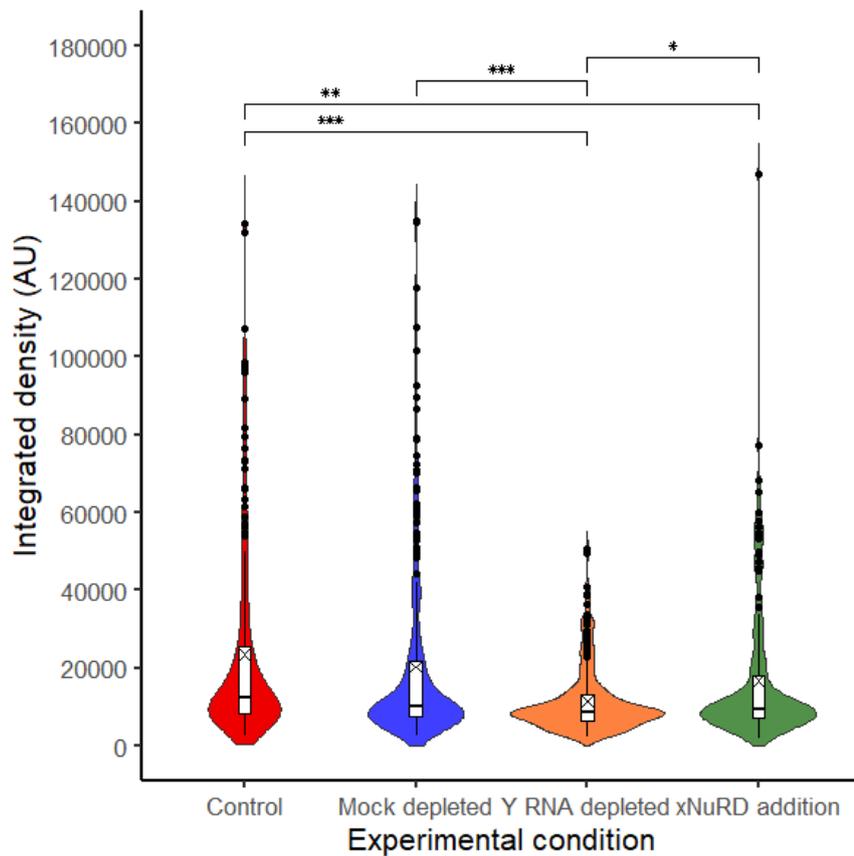
**Figure 7.1:** A schematic of ds-eloSeq method. The potential replication origin sites (a) undertook DNA replication initiation and elongation in a 3 hour replication reaction, in which replication initiation and limited elongation took place in the first 15 minutes (as seen in ds-*ini*Seq) (b) and further elongation and potential origin firing took place in the next 2 hours and 45 minutes (c), under standard conditions (synchronised G1 phase nuclei were incubated with cytosol from proliferating cells and a mixture of dNTPs, NTPs and an ATP regenerator). The newly synthesised DNA incorporated the heavy BrdU (b&c). Total genomic DNA was fragmented to produce a mixture of replicated heavy-light (HL) and unreplicated light-light (LL) DNA (d), which was separated on a  $\text{Cs}_2\text{SO}_4$  density gradient (e). Purified HL and LL DNA, which was separated from replication intermediates, DNA repair fragments (f) and RNA (RI = 1.3814 AU), underwent Illumina sequencing and mapped to the human hg38 genome (g).

The primed and licensed human replication origins in the late G1-phase template nuclei (Fig.7.1a) fired within the first 15-minutes of the *in vitro* replication reaction (Fig.7.1b). Replication continued during the next 2 hours and 45-minutes of the *in vitro* replication reaction (resulting a total reaction time of 3-hours), where elongation and firing of origins that occur later in replication took place (Fig.7.1c). In this reaction, the light dTTPs were entirely replaced with the heavy nucleotide, BrdU, which was incorporated into the newly synthesised DNA (Fig.4.1b/c).

The purification, fragmentation and separation of the replicated HL and unreplicated LL DNA were the same as that of ds-iniSeq. Total genomic DNA was treated with proteinase K and sonicated into 100-1000bp fragments (Fig.7.1d). The replicated HL and unreplicated LL DNA were separated via density equilibrium centrifugation using a caesium sulphate gradient (Fig.7.1e), where HL and LL DNA were present in fractions at refractive indices of 1.3700 and 1.3675 respectively and RNA was removed in a fraction at a RI of around 1.3814 (2). LL DNA was isolated from the first density gradient and HL DNA was isolated from the second. Pools of purified and separated HL and LL DNA were desalted and underwent Illumina sequencing (Fig.7.1f). The sequencing data was mapped and aligned to the hg38 genome (Fig.4.3g). For these ds-eloSeq experiments, Illumina sequencing was performed by our collaborators at the NGS facility at the Francis Crick Institute (library preparation by R. Jones, J. Smith group) and mapping and alignment was performed by TK.

As with ds-iniSeq, the replication reaction of ds-eloSeq can be manipulated to assess the impact of various replication factors on the extent replication elongation; here, I examined the hY RNAs and xNuRD.

In parallel with the ds-eloSeq reactions, I quantified DNA synthesis by immunofluorescence microscopy in the human cell-free system for the same experimental conditions. For these reactions, digoxigenin tagged dUTP replaced the BrdUTP and was incorporated into newly synthesised DNA. The nuclei were then spun onto glass slides, treated with propidium iodide (for detecting DNA) and anti-digoxigenin fluorescein antibodies (for detecting digoxigenin incorporation), and imaged using confocal microscopy. Normally these images are used to determine the percentage of replicating nuclei, as shown in Fig.6.1. For the experiments carried out alongside the ds-eloSeq reactions, I quantified the extent of fluorescein incorporation on a per-nucleus basis, and therefore digoxigenin-tagged newly synthesised DNA in the imaged replicating nuclei using a Fiji script generated by GG (Fig.7.2).



**Figure 7.2:** The human cell-free experiments that were conducted alongside the replicate 1 ds-eloSeq experiments. In the human cell-free system experiment, template nuclei, isolated from mimosine treated EJ30 cells (synchronised to late G1), were incubated (3 hours; 37°C) in the presence of a physiological replication buffer, NTPs, dNTPs (including digoxigenin-dUTP), creatine kinase and: with cytosol (proliferating HeLa cells) (“Control”); with cytosol (proliferating HeLa cells) pre-incubated with an oligonucleotide complementary for the bacteriophage T3 RNA, to act as a control for oligonucleotide/RNase H degradation of RNAs (“Mock depleted”); with cytosol (proliferating HeLa cells) pre-incubated with an oligonucleotide complementary for hY RNAs, to degrade/eliminate hY RNAs (“Y RNA depletion”); and with cytosol (proliferating HeLa cells) pre-incubated with an oligonucleotide complementary for hY RNAs, to degrade/eliminate hY RNAs, and the addition of partially purified xNuRD (“xNuRD addition”). Following the replication reaction incubation, the nuclei for each condition were transferred to individual microscope slide and stained. The nuclei were imaged using confocal microscopy and the images were analysed using a FIJI script, written by G.Guilbaud (MRC-LMB, Cambridge), to quantify the incorporation of digoxigenin-dUTPs in nuclei using the measure of integrated density (sum of pixel intensities of the regions of interest; in this case, each nucleus). The violin plot shows the integrated densities of the nuclei imaged in each experimental condition. An ANOVA and subsequent Tukey’s post-hoc test were performed to assess significance; the Tukey’s test results are shown on the plot and \*\*\* indicates  $p < 0.001$ , \*\* indicates  $p < 0.01$  and \* indicates  $p < 0.05$ ; all other tests were not significant.

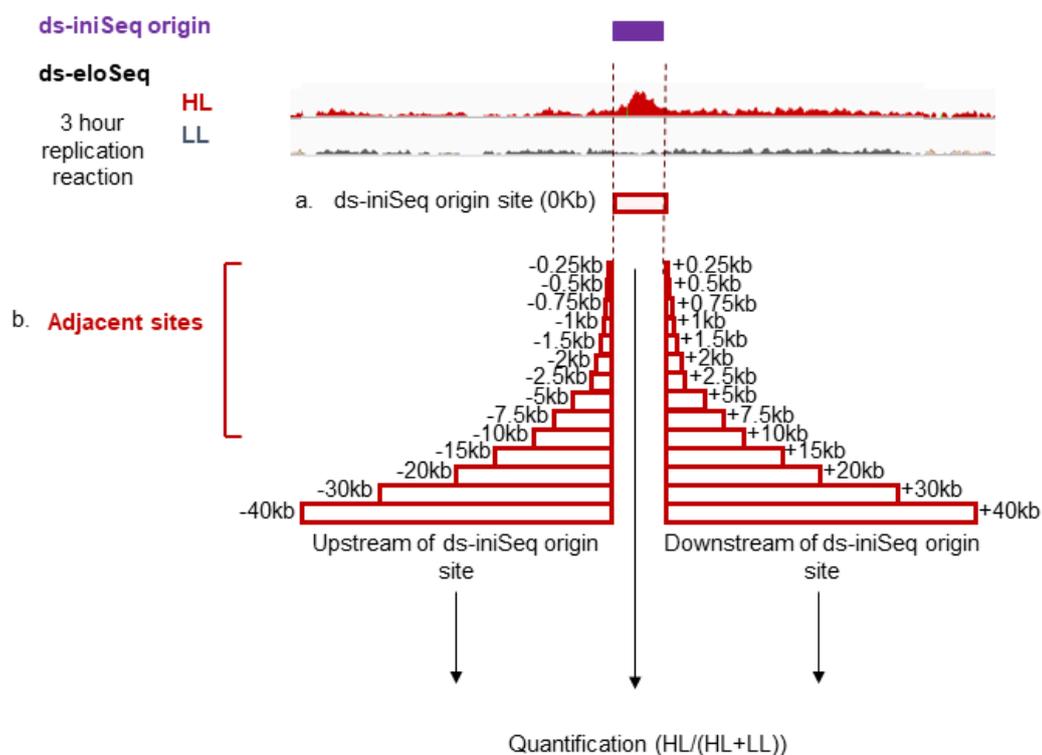
The integrated pixel densities (which is the sum of pixel intensities of the region of interest; in this case each nucleus) of the fluorescein incorporation in the nuclei of the control condition was not significantly higher than that of the mock-depleted condition (Fig.7.2). The integrated densities of the nuclei in the Y RNA-depleted condition were significantly lower than nuclei in the control and mock-depleted conditions. The integrated densities of the mock-depleted condition nuclei were not significantly different to that of the xNuRD condition. The integrated densities of the nuclei in the xNuRD addition condition were also significantly greater than that of the Y RNA-depleted. Additionally, the range of integrated densities of nuclei in the

control, mock-depleted and xNuRD addition conditions were much greater than that of the Y RNA condition.

The extent of fluorescein incorporation in these nuclei (as assessed by integrated density) indicated the amount of replication taking place. This showed that more replication took place in the nuclei of the control, mock-depleted and xNuRD addition conditions than in the Y RNA-depleted nuclei. This concurred with previous findings (3)(Fig.6.1) and together confirmed that at the replication foci level, Y RNAs were required for DNA replication and xNuRD could functionally substitute for hY RNAs.

### 7.2.1 Preliminary analysis of elongation using ds-eloSeq data

I devised a preliminary method of assessing the extent of activation of potential origins in the ds-eloSeq data and elongation away from them (Fig.7.3).



**Figure 7.3:** A schematic of the preliminary method employed to assess the extent of replication elongation away from the origins defined by the ds-iniSeq data. (a) The extent of replication as measured by the ratio of the relative of sequenced HL DNA to total DNA (HL/(HL+LL)) was assessed, in the ds-eloSeq data, at the locations of the previously identified by ds-iniSeq. (b) The site activity (HL/(HL+LL)) of sites adjacent to the previously identified ds-iniSeq origin locations were calculated from the ds-eloSeq sample, which determined the extent of replication elongation away from these sites. Site activity was measured 0.25Kb, 0.5Kb, 0.75Kb, 1Kb, 1.5Kb, 2Kb, 2.5Kb, 5Kb, 7.5Kb, 10Kb, 15Kb, 20Kb, 30Kb and 40Kb upstream and downstream of the ds-iniSeq origin locations in the ds-eloSeq samples. The read count of sequenced HL DNA and LL DNA were determined using the feature probe generator and read count quantification tools in SeqMonk.

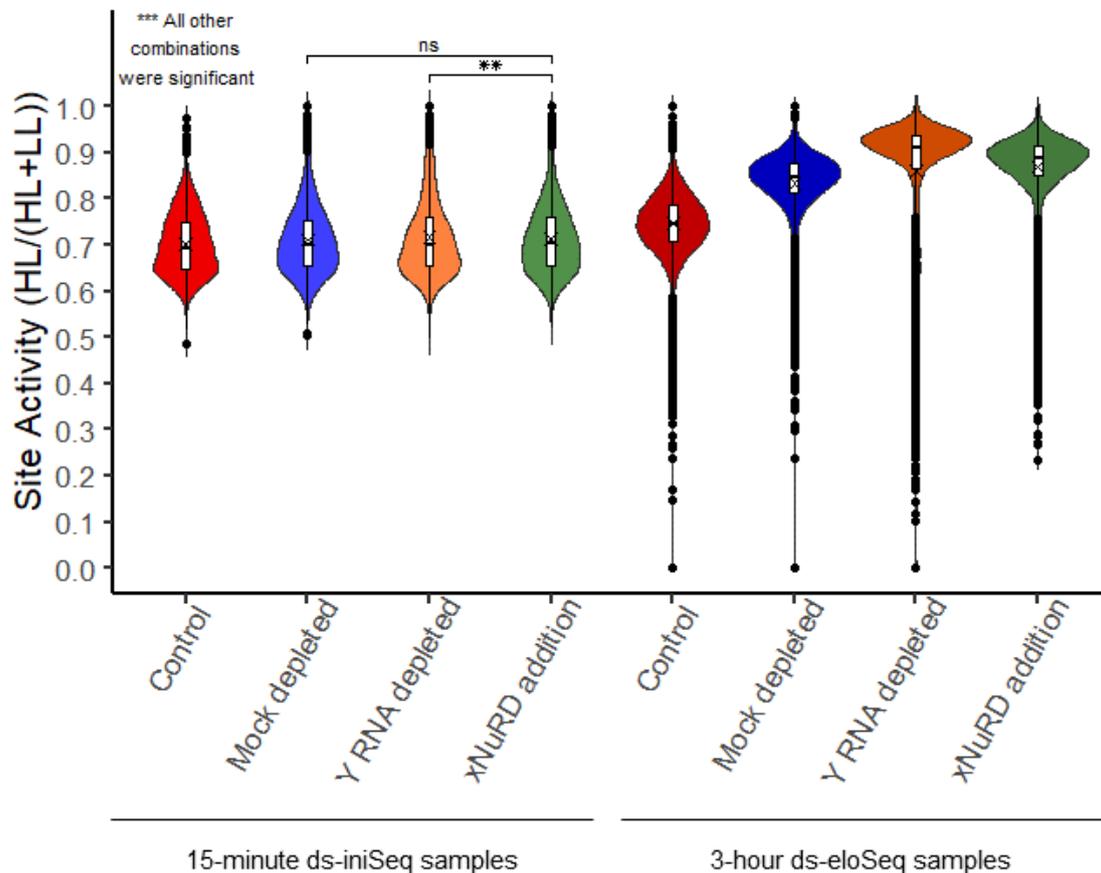
I used the ds-iniSeq origins (replicate 1 origins present in replicates 2 and 3) identified in chapters 5/6 to provide the locations of origin sites for analysis in ds-eloSeq. I determined the relative replication activity levels of these sites, in the sequenced ds-eloSeq data files, from

the ratio of the relative amount of sequenced (ie read count) HL DNA to total DNA ( $HL/(HL+LL)$ ), as previously done in ds-iniSeq (Fig.7.3a). This allowed me to interrogate the change in activity of potential origin sites at a later point of the DNA replication reaction *in vitro* (ie following a 3-hour rather than a 15-minute incubation).

To assess the extent of replication elongation away from replication origins, I measured the relative replication activity ( $HL/(HL+LL)$ ) of sites reaching to 0.25Kb, 0.5Kb, 0.75Kb, 1Kb, 1.5Kb, 2Kb, 5Kb, 7.5Kb, 10Kb, 15Kb, 20Kb, 30Kb and 40Kb upstream and downstream of the ds-iniSeq origin locations (Fig.7.3b) (ie the sites increased in size). I also assessed the replication activity ( $HL/(HL+LL)$ ) of the same sites up and downstream in the 15-minute ds-iniSeq data for comparison. These analyses were conducted using SeqMonk's feature probe and read count quantitation tools (4). Together these data demonstrate the difference in the extent of replication that took place over the first 15-minutes and 3-hours of a replication reaction.

### 7.2.2 Ds-iniSeq origin sites

I compared the relative origin activities of the 14,126 control, 10,357 mock-depleted, 7,076 Y RNA-depleted and 10,329 xNuRD addition ds-iniSeq origins (after a 15-minute replication reaction) to the activities of sites at the same locations and sizes in the corresponding ds-eloSeq experimental conditions (3-hour reaction reaction) (Fig.7.4).



**Figure 7.4:** The relative activities (HL/(HL+LL)) of the ds-iniSeq origins (replicate 1 origins also present in replicates 2 and 3) in each experimental condition from the experiments: control; mock-depleted; Y RNA-depleted; and xNuRD addition. The relative activities of the 14,126 control, 10,357 mock-depleted, 7,076 Y RNA-depleted and 10,329 xNuRD addition ds-iniSeq origins locations, in the corresponding ds-eloSeq samples (replicate 1). An ANOVA and subsequent Tukey's post-hoc test were performed to assess significance; the Tukey's test results are shown on the plot "ns" indicates a not significant result and \*\* indicates  $p < 0.01$ ; all other tests were highly significant \*\*\* where  $p < 0.001$ .

The activities of the control, mock-depleted, Y RNA-depleted and xNuRD addition ds-iniSeq origins (15-minute replication reaction) were significantly different from one another, except for the mock-depleted and xNuRD addition ds-iniSeq origins which were similar (Fig.7.4). In chapter 6, I determined that the Y RNA-depleted ds-iniSeq origins contained a subgroup of origins with unexpectedly high activities that were Y RNA-independent and that shifted the overall activities upwards. This may explain the lack of reduction in origin activity when compared to that of the other experimental conditions (Fig.4.4).

The same 15-minute ds-iniSeq origin locations in the 3-hour ds-eloSeq samples were used to measure the activities at those sites and showed that the relative activity of each experimental condition was significantly higher than for the corresponding 15-minute ds-iniSeq origins (Fig.7.4), suggesting that there was a greater extent of origin firing after a 3-hour replication reaction compared to 15-minutes. The ranges of these activities were much larger than the ranges of the corresponding ds-iniSeq origins. The lower activities of the few sites in the ds-eloSeq samples when compared to the ds-iniSeq origins indicated that there was less replication taking place at these sites in the 3-hour reaction, which may result from

these origins not having a consistent probability of firing and may have remained dormant in the majority of nuclei in the ds-eloSeq samples. However, the interquartile ranges were much smaller and higher for all conditions in the ds-eloSeq samples than that of the corresponding ds-iniSeq origins. In the 3-hour ds-eloSeq data, the Y RNA-depleted sites possessed the highest activities, followed by the xNuRD addition sites, followed by the mock-depleted sites whilst the control sites possessed the lowest activities. These suggest that, surprisingly, Y RNA removal from the 3-hour replication reactions of ds-eloSeq, resulted in an increase in replication activities identified as replication origins by ds-iniSeq during later incubation times after the 15 minutes used in the ds-iniSeq samples.

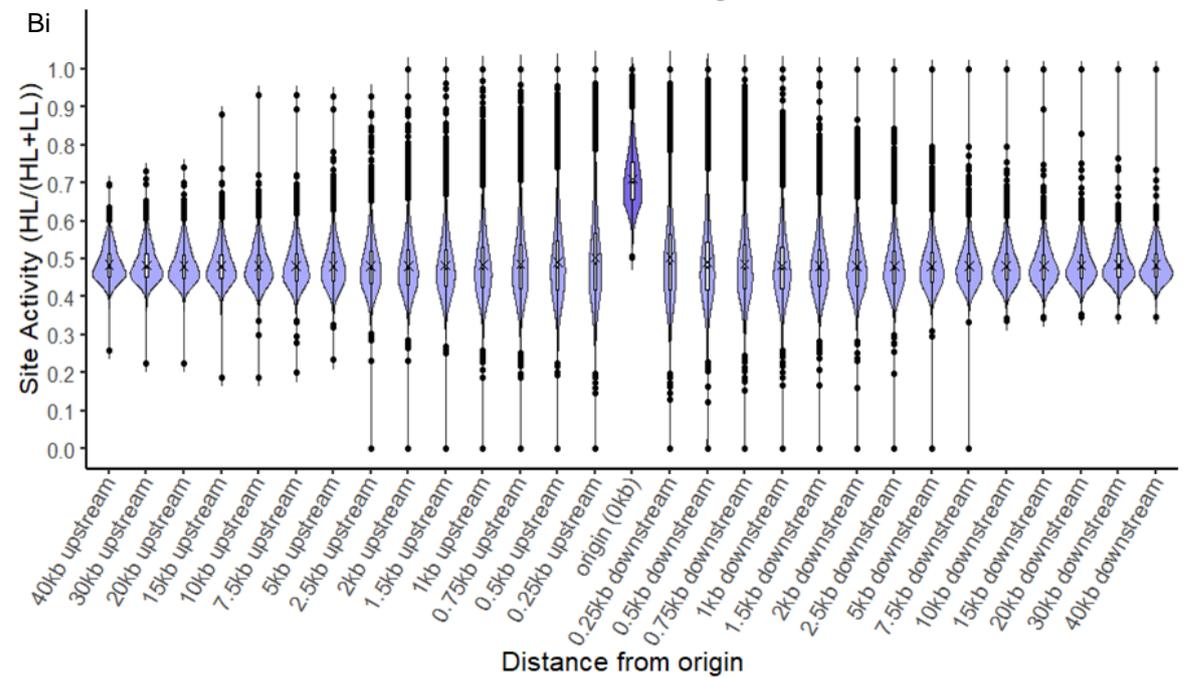
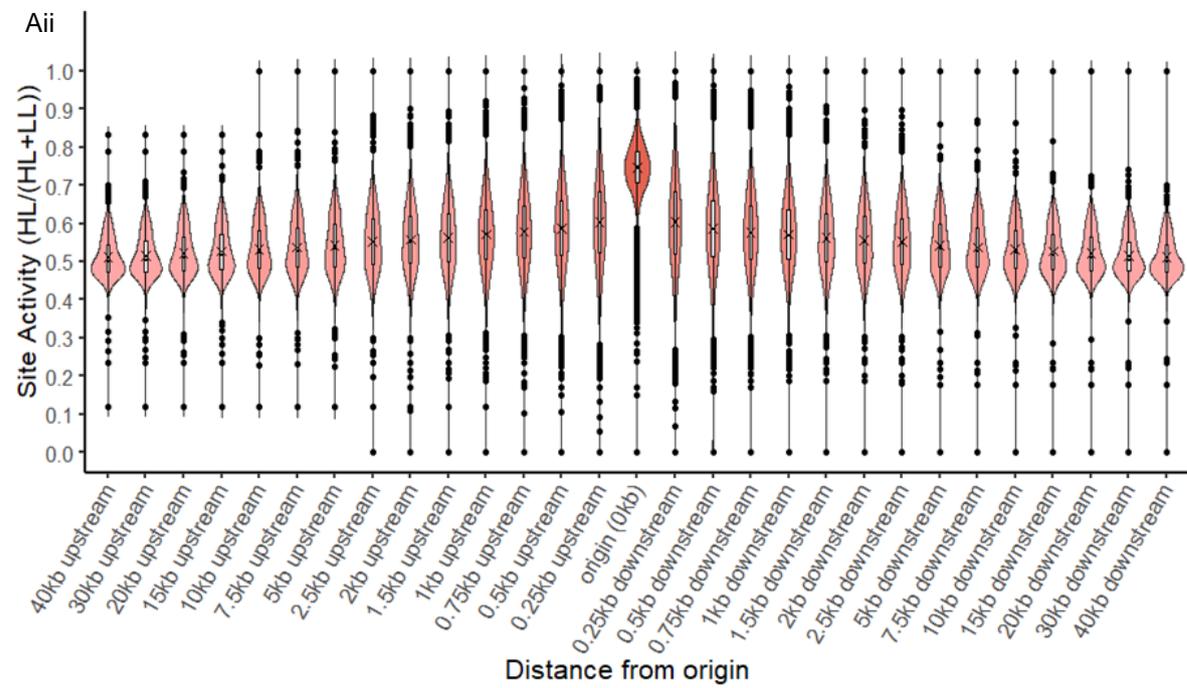
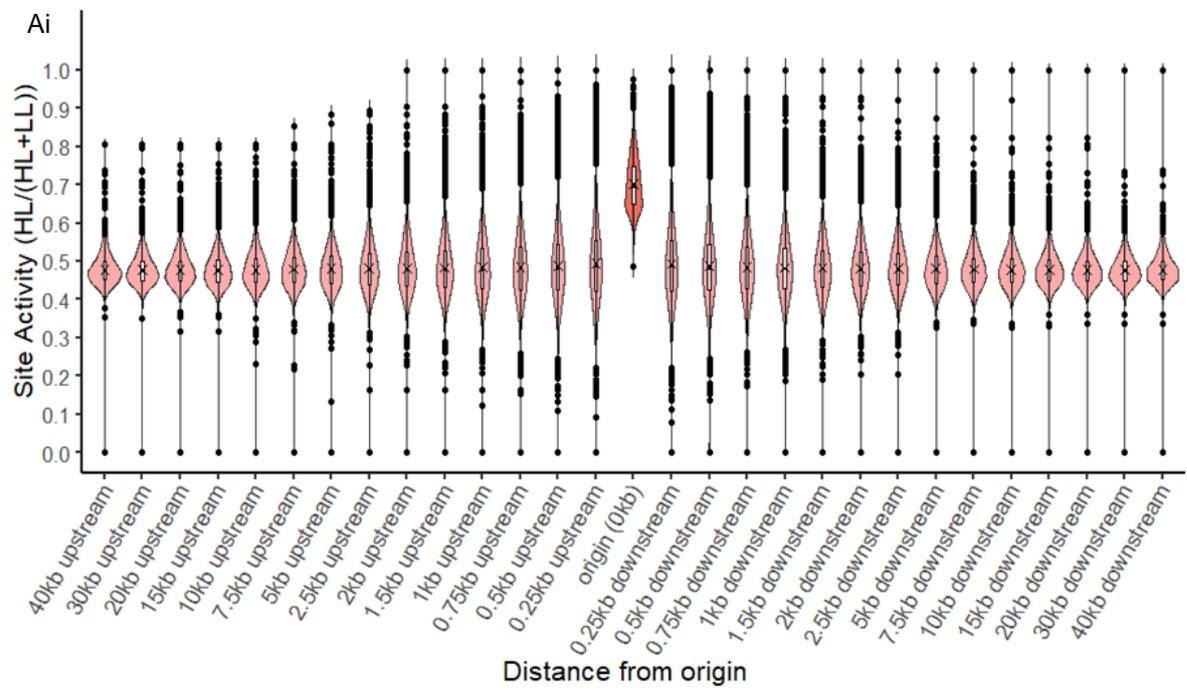
This was the inverse of the observations in the same experimental conditions in the human cell-free system, where nuclei in the control conditions have the highest measured levels of replication, closely followed by nuclei in the mock-depleted condition. Y RNA removal dramatically reduced replication levels and xNuRD restored them to similar levels to the mock-depleted condition (Fig.6.1/7.2) (3). The difference between both observations may be explained by the differences in resolution of both methods. The human cell-free system assesses DNA replication at the resolution level of replication foci (containing 0.5-1Mb, encompassing multiple origins and elongating replication forks), whereas the ds-iniSeq/ds-eloSeq analysis is at the resolution level of short replication origins.

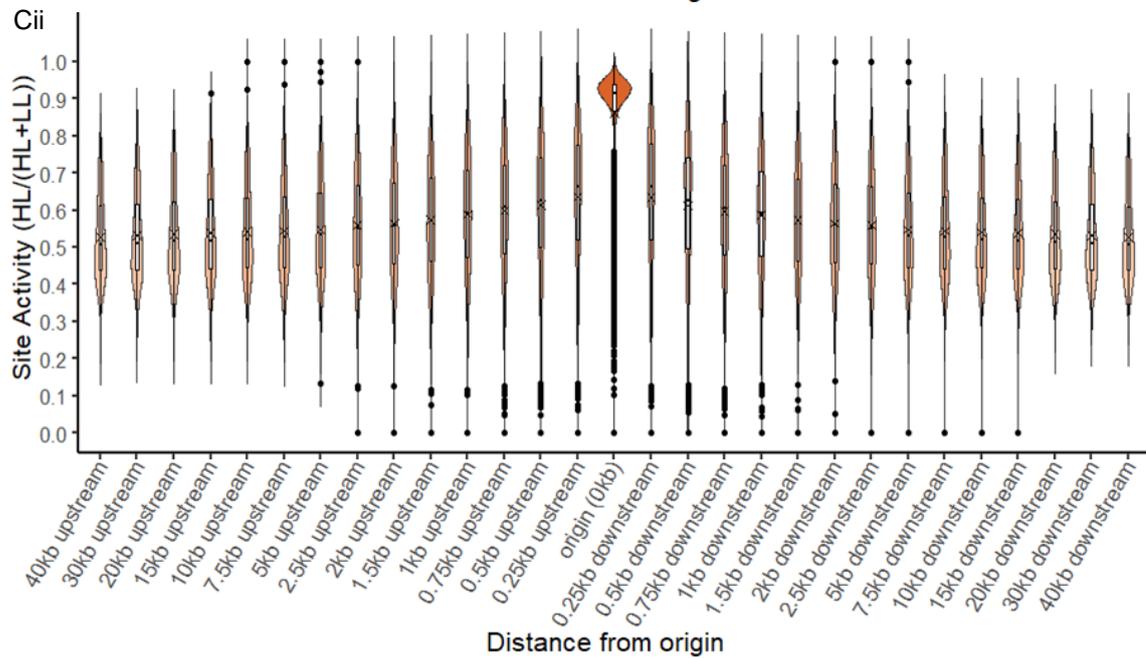
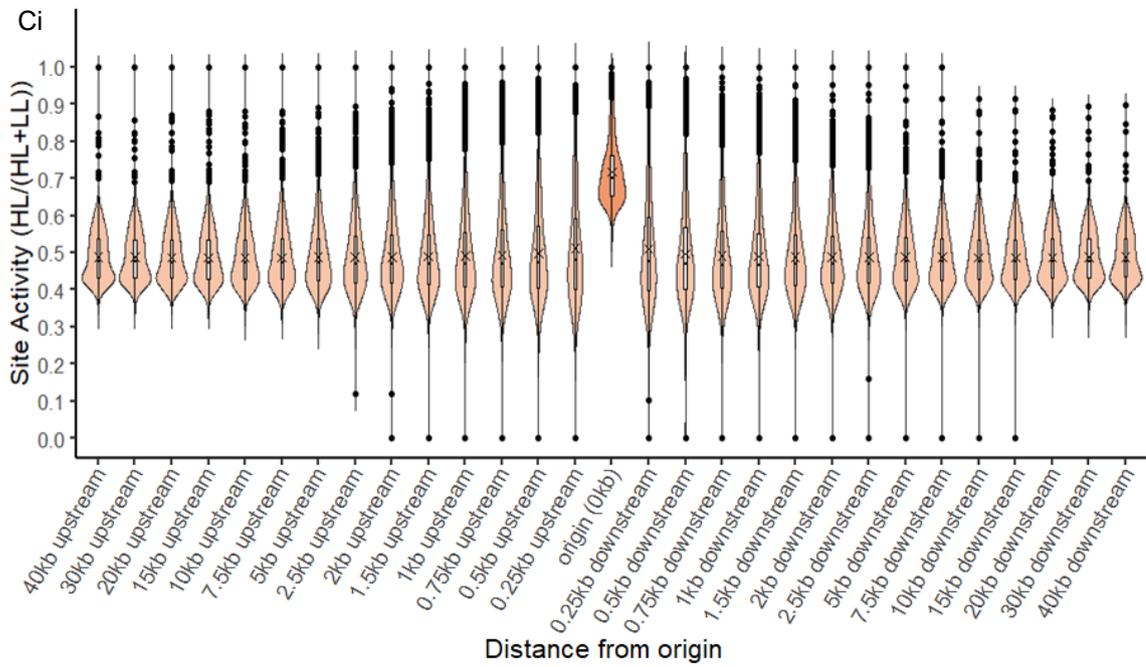
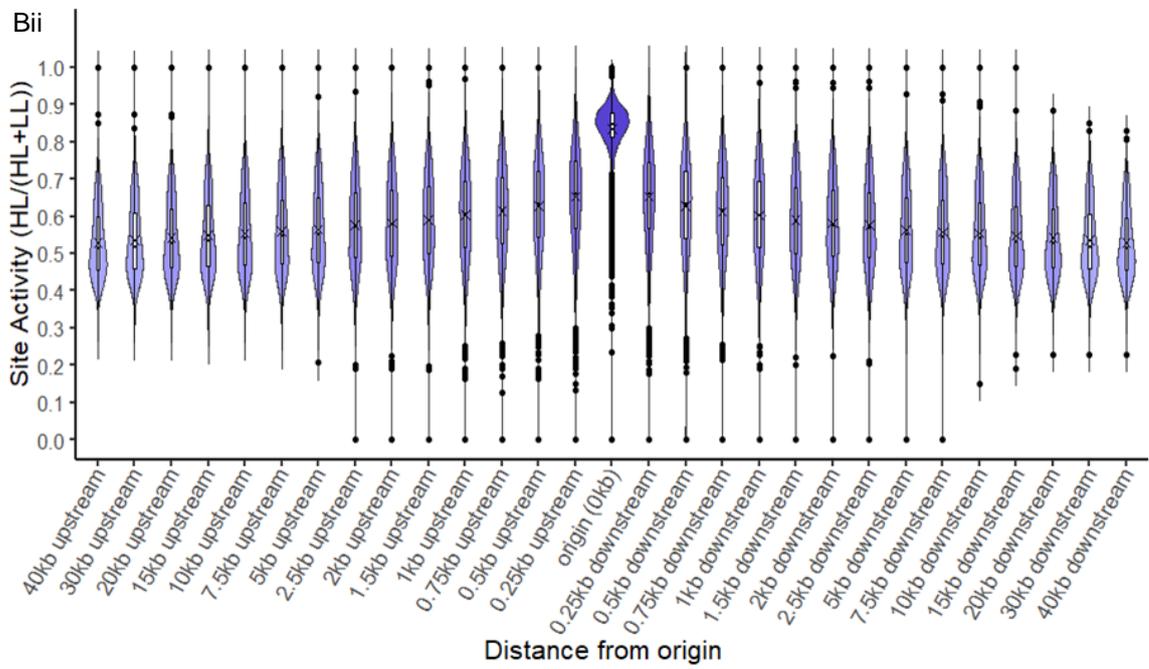
Together, these data showed that the depletion of Y RNAs was associated with an increase in the origin site activity after a 3-hour replication reaction but with decreased DNA replication at the replication foci level. This may indicate a potential role for Y RNAs in DNA replication elongation. However, it has been shown that Y RNAs are not required for replication elongation and do not associate with replication fork proteins (5,6). In chapter 6, I demonstrated that Y RNA removal resulted in lower origin activities in the 15-minute ds-iniSeq origins that were dependent on Y RNAs (Fig.6.4C). Therefore, a role for Y RNAs in elongation/fork progression is unlikely and may instead suggest a role for Y RNAs in the transition of origin firing (initiation) to replication fork progression (elongation).

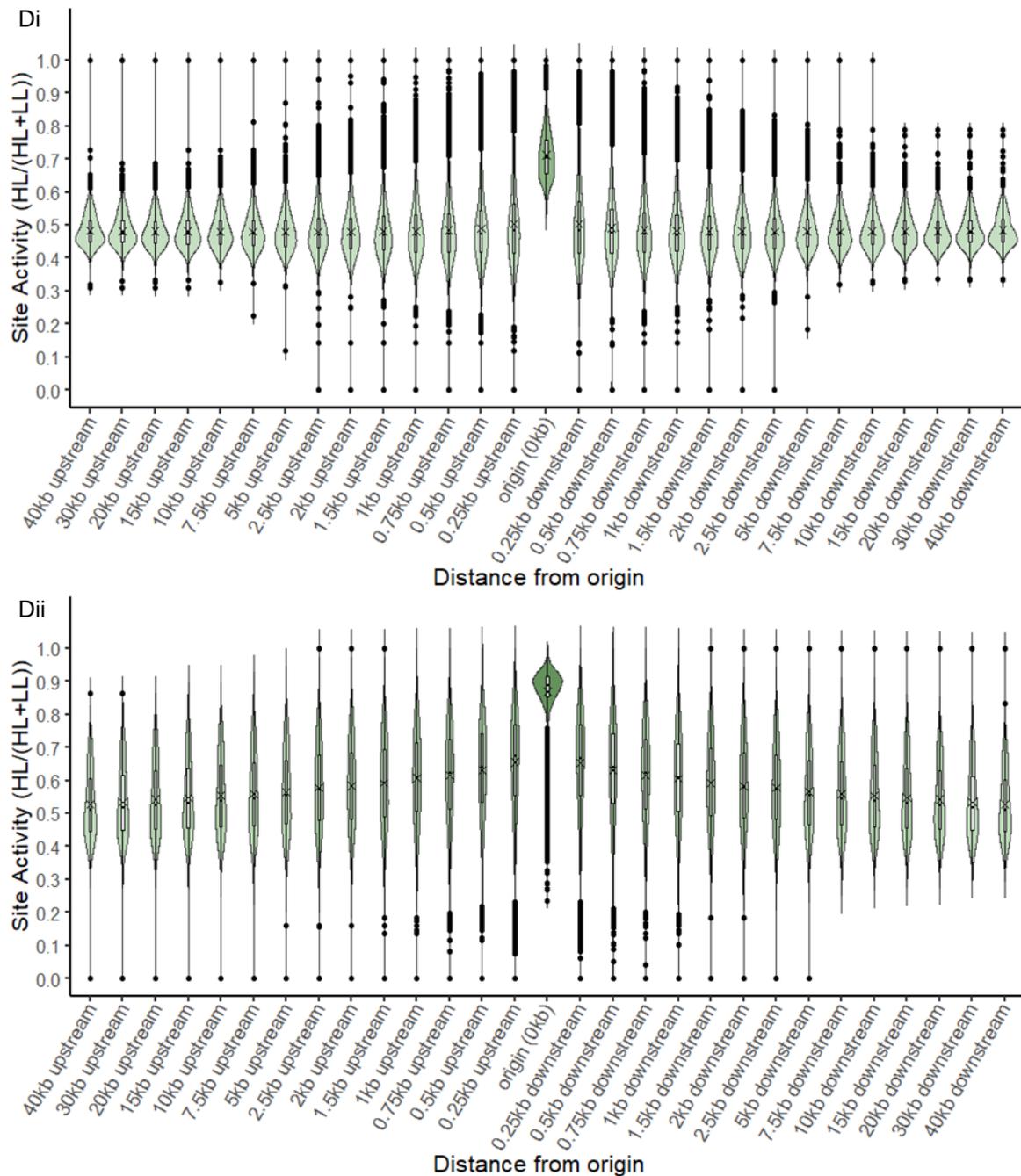
Alternatively, the Y RNA-independent origins identified in the 15-minute ds-iniSeq samples (Fig.6.4C) may become even more active in a 3-hour reaction. There were lower numbers of the Y RNA-dependent origins in the Y RNA-depleted experimental condition compared to the mock-depleted and xNuRD addition conditions and these Y RNA-depleted origins may be compensating for the reduced number of origins by firing more efficiently later in replication. It has been suggested that the number of active replication forks may influence replication fork progression; *mec1-100* cells show both firing of excess origins and slower replication fork progression (7). This relationship may be applicable to the observations shown here.

### *7.2.3 Elongation away from ds-*ini*Seq origin sites*

I determined the activity levels ( $HL/(HL+LL)$ ) immediately upstream and downstream of the ds-*ini*Seq origin sites locations in all experimental conditions, as described in Fig.7.3. I removed any activity values if the upstream and downstream sites ran into an adjacent origin. I plotted all the replication activities of the origin sites and the corresponding sites upstream and downstream, on a violin plot, for all four experimental conditions in the ds-*elo*Seq data (3-hour replication reaction) and the ds-*ini*Seq data (15-minute replication reaction) for comparison (Fig.7.5).







**Figure 7.5:** Violin plots of the relative activities of the ds-iniSeq origins and the adjacent sites of the 15-minute ds-iniSeq samples for the control (Ai), mock-depleted (Bi), Y RNA-depleted (Ci) and xNuRD addition (Di) experimental conditions. Violin plots of the activities of the ds-iniSeq origin locations and the adjacent sites of the 3-hour ds-eloSeq (replicate 1) samples for the control (Aii), mock-depleted (Bii), Y RNA-depleted (Cii) and xNuRD addition (Dii) experimental conditions. The sites were defined in accordance with the schematic in figure 7.3.

The violin plots of the relative activities of origins and adjacent sites upstream and downstream of the 15-minute ds-iniSeq samples showed that in all experimental conditions (Fig.7.5Ai/Bi/Ci/Di), that there was no difference in overall replication occurring at sites adjacent to the origins, irrespective of the activity levels of the origins.

In all experimental conditions, the activity levels of the sites adjacent to the origins increased in the 3-hour ds-eloSeq samples (Fig.7.5Aii/Bii/Cii/Dii) compared to the corresponding sites

in the 15-minute ds-*iniSeq* samples. This suggests that there was additional replication elongation and replication fork progression away from their initiation sites in the 3-hour ds-*eloSeq* replication reactions compared to the 15-minute ds-*iniSeq* replication reactions.

In all experimental conditions, the violin plots of the relative activities of origins and adjacent sites upstream and downstream of the 3-hour ds-*eloSeq* samples (Fig.7.5Aii/Bii/Cii/Dii) showed that sites adjacent to the origins possessed almost identical activity levels in sites upstream, at all distances from the origins, compared to the corresponding downstream sites. These suggest that there was no bias for replication fork progression upstream or downstream and that overall replication elongation/fork progression occurred in a bi-directional fashion in all experimental conditions; and that any potential unidirectional replication fork progression from an origin were equally distributed.

The activities of the sites adjacent to the origins in the mock depletion ds-*eloSeq* sample (Fig.7.5Bii) were slightly higher than that of the Y RNA depletion ds-*eloSeq* samples (Fig.7.5Cii), at all distances upstream and downstream. The activities of the origin sites in the mock-depleted ds-*eloSeq* samples were much lower than that of the Y RNA-depleted ds-*eloSeq* samples. Therefore, the difference between the activities of the sites upstream and downstream of the origins and the activities of the origins was greater in the Y RNA-depleted condition compared to the mock-depleted. This indicates that the activity of an origin site does not necessarily correspond/relate to the activity levels of elongation/replication fork progression away from those origins.

Additionally, these data show that Y RNA removal was associated with increased origin site activity, but this was not the case for replication elongation into sites adjacent to the origins, which offers a potential explanation for why Y RNA removal in the human cell-free system resulted in a reduction in observed replication activity.

The activities of the sites adjacent to the origins in the xNuRD addition ds-*eloSeq* sample (Fig.7.5Dii) were almost identical to that of the mock depletion ds-*eloSeq* samples (Fig.7.5Bi), at all distances upstream and downstream. However, the difference between the activities of the sites upstream and downstream of the origins and the activities of the origins of the xNuRD addition condition (Fig.7.5Dii) was highly similar to that of the Y RNA-depleted condition (Fig.7.5Cii), at all distances upstream and downstream. These suggest that xNuRD did not restore the relationship between origin firing and replication elongation to similar levels as the mock-depleted condition, which contradicts the published data using immunofluorescence microscopy showing that xNuRD does substitute for Y RNAs in DNA replication (3).

I show here that in the 3-hour ds-*eloSeq in vitro* replication reaction, Y RNA depletion increased origin firing but interfered with elongation away from these origin sites. This is

consistent with the published data using immunofluorescence microscopy to assess DNA replication in template nuclei of a 3-hour *in vitro* replication reaction. I determined in chapter 6 that Y RNA depletion resulted in a biphasic response; all origins fired less efficiently and consequently fewer origins were called. The observations of origin site activity are inconsistent with the data presented here (chapter 6) I am not able to determine whether the number of activated origins in the 3-hour ds-eloSeq data is affected by Y RNA depletion.

To resolve this, a new method for calling origins in the 3-hour ds-eloSeq samples is required, as MACS is unable to distinguish between replication elongation and replication origins. I present here preliminary work on the development of the Wilkes-Mookerjee (WM) method for calling origins in the 3-hour ds-eloSeq samples in section 7.3. Once the WM method is fully developed, it could be applied to the mock-depleted, Y RNA-depleted and xNuRD addition ds-eloSeq data, to ascertain the impact on of Y RNAs and xNuRD on origins that fire later in S-phase.

In chapter 6, I also showed that a smaller number of extra origins became activated upon Y RNA depletion and were unique to this experimental condition. These origins were highly active, present in later replication timing windows and showed a substantially reduced colocalisation with genomic and epigenetic features normally associated with higher origin activities in the presence of Y RNAs. It is possible that these Y RNA-independent origins (that were included in the origins interrogated in the Y RNA-depleted ds-eloSeq data), may have introduced a bias/skew for higher activities in these data. Additionally, they may be a subset of origins of an inherently different nature to the ds-iniSeq origins affected by Y RNAs. And, consequently, may have contributed to the lower activities of the sites adjacent to the origin sites.

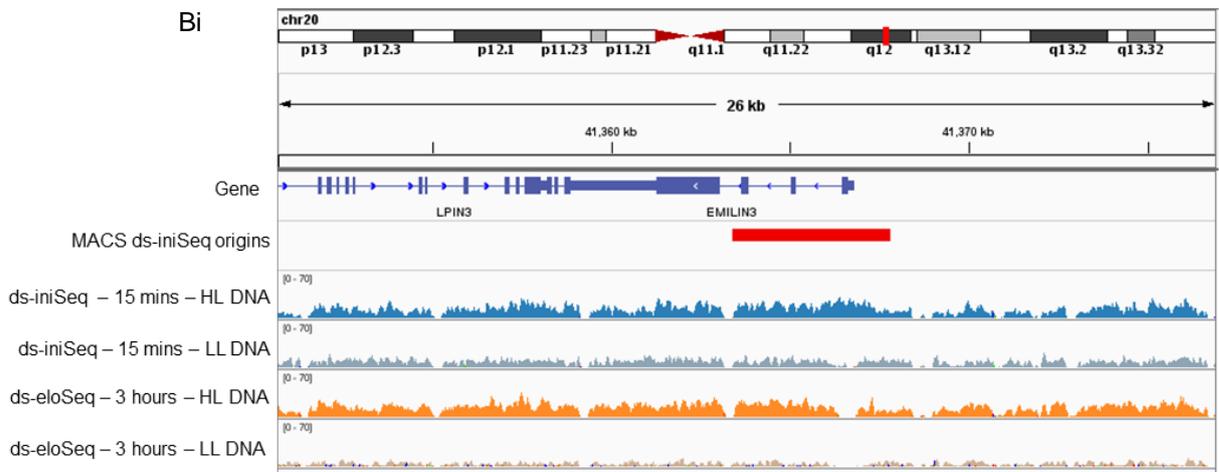
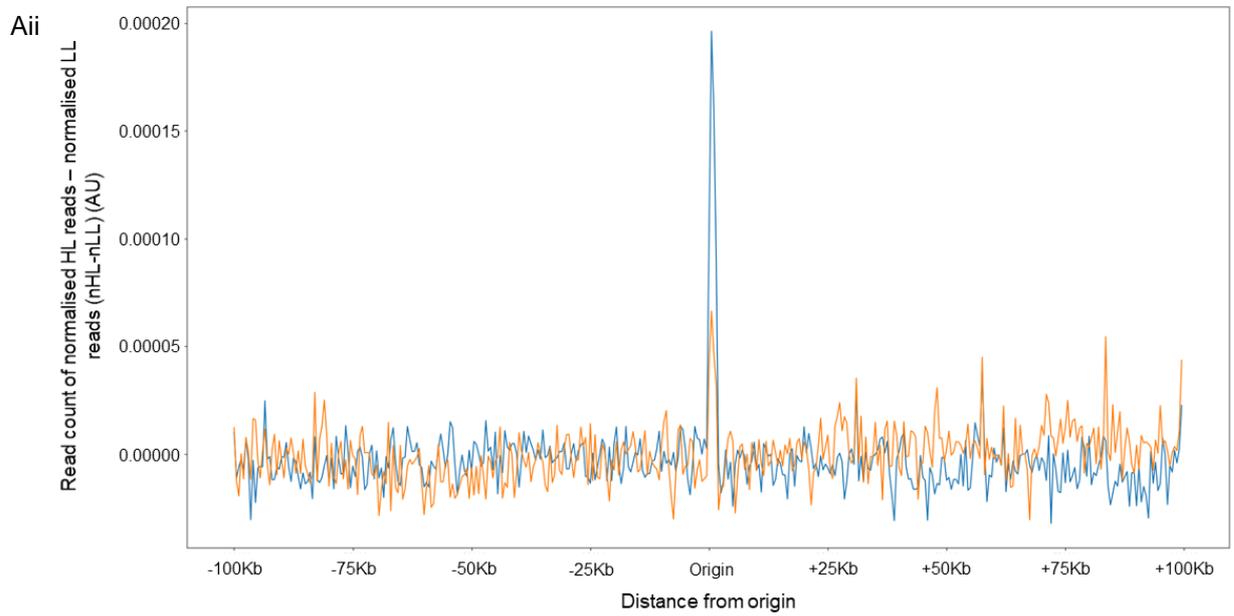
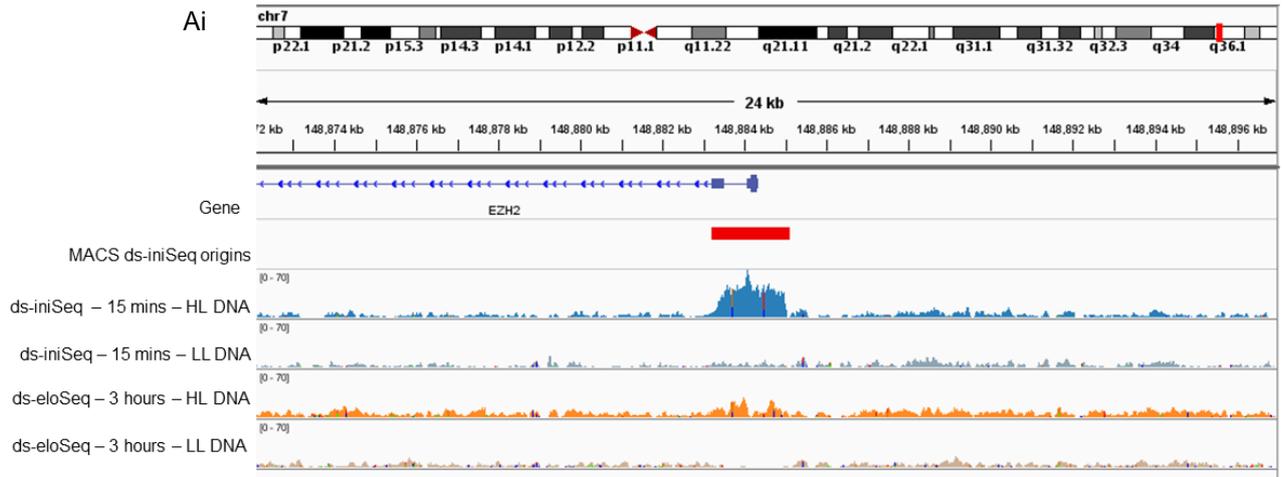
In an attempt to reconcile my findings in the ds-iniSeq data and the ds-eloSeq data, I hypothesised that Y RNAs may be involved in the firing of origins and the transition of initiation to elongation. From the data presented in chapters 5 and 6, I proposed that Y RNAs interacted with chromatin remodellers to bring about the generation of an epigenetic environment conducive with replication initiation. This environment may also be required for the transition from origin firing to replication fork progression in elongation. Potentially, when this transition cannot take place in the absence of Y RNAs, origin firing slowly accumulates, resulting in the higher replication activities at origin sites in the ds-eloSeq samples. Clearly this conjecture requires a great deal of future work to establish its validity.

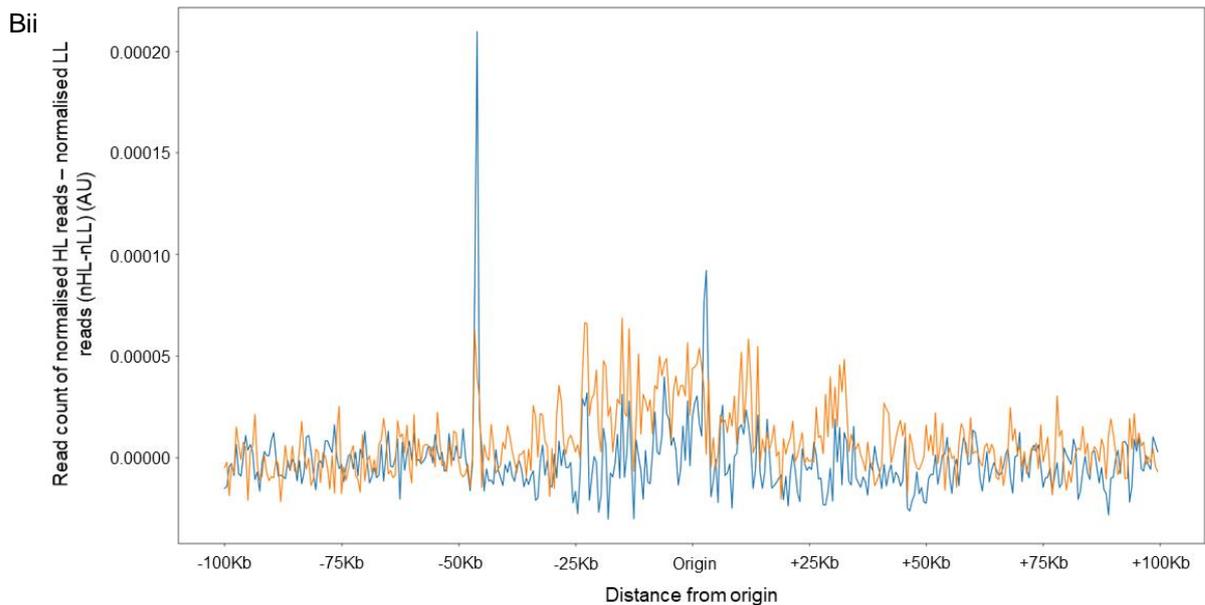
## 7.3 Results and Discussion – Development of origin calling method in ds-eloSeq data

### 7.3.1 “Wilkes-Mookerjee” (WM) method for origin calling in 3-hour ds-eloSeq data

MACS peak calling was designed and optimised for short sharp peaks (1) which makes it inappropriate for calling origin peaks in the 3-hour ds-eloSeq data as replication elongation took place in addition to origin firing. With my colleague S.Mookerjee (St Catharine’s College, Cambridge) I formulated a multidisciplinary approach to develop a computational method to call replication origin sites in the 3-hour ds-eloSeq sample files. Development of this method, termed “Wilkes-Mookerjee” (WM), is ongoing and may require future refinement, but in its current form, has successfully identified potential origin sites in both the ds-iniSeq and ds-eloSeq data files. The identification/calling of origin sites in the 3-hour elo-Seq data is necessary in order, to subsequently identify sites of elongation away from the origins.

We verified the WM method using the control ds-iniSeq and control ds-eloSeq sample files from replicate 1. The HL and LL sample files were normalised to the total read number in each file, so they were directly comparable. The normalised read count of the LL DNA files (which constituted the background) was subtracted from the normalised read count of the HL DNA files (which constituted the background + signal) in 500bp windows, resulting in an overall normalised read count of the signal only ( $n_{HL} - n_{LL}$ ); examples are shown in Fig.7.6.





**Figure 7.6:** (Ai) The IGV image of the EZH2 promoter origin site. The ds-inoSeq origins (replicate 1 origins found in replicates 2 and 3) called by MACS peak caller are in red. The replicate 1 HL and LL DNA files of the 15-minute ds-inoSeq (blue/grey) and the 3-hour ds-eloSeq samples (orange/grey) were shown. (Bi) The IGV image of the EMILIN3 origin site. The ds-inoSeq origins (replicate 1 origins found in replicates 2 and 3) called by MACS peak caller are in red. The HL and LL DNA files of the 15-minute ds-inoSeq (blue/grey) and the 3-hour ds-eloSeq (orange/grey) samples were shown. (ii) The HL and LL sample files were normalised to the total read number of each file and the normalised LL DNA files was subtracted from the corresponding normalised HL DNA files in 500bp windows in the 15-minute ds-inoSeq (blue) and 3-hour ds-eloSeq (orange) sample. These were visualised for the origins identified at the EZH2 promoter (Aii) and the EMILIN3 gene (Bii).

The IGV image showed the HL DNA and LL DNA of replicate 1 ds-inoSeq and ds-eloSeq control reactions, at the EZH2 promoter origin (Fig.7.6Ai). The ds-inoSeq origin identified by MACS are shown in red, and the ds-inoSeq HL and LL DNA sample files (blue) revealed a strong enrichment of HL DNA compared to the LL DNA. Whereas, the ds-eloSeq HL and LL DNA sample files (orange) showed a smaller HL DNA enrichment at the same EZH2 site but a depletion of LL DNA at the corresponding area.

The nHL–nLL profile of the ds-inoSeq samples (blue) and the ds-eloSeq samples (orange) were plotted at the EZH2 promoter origin +/-100kb (Fig.7.6Aii). These showed a large peak in the ds-inoSeq sample and a smaller but substantial peak in the ds-eloSeq sample at the identified origin at the EZH2 promoter, which reflected the observations in the IGV images. Additionally, there was fluctuation at the background level around 0.

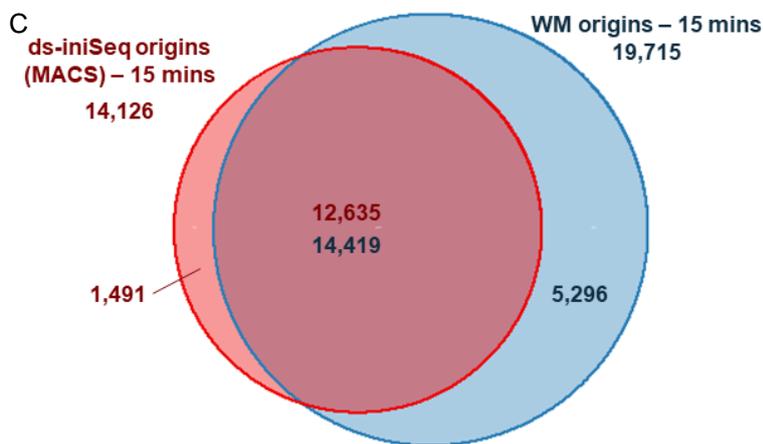
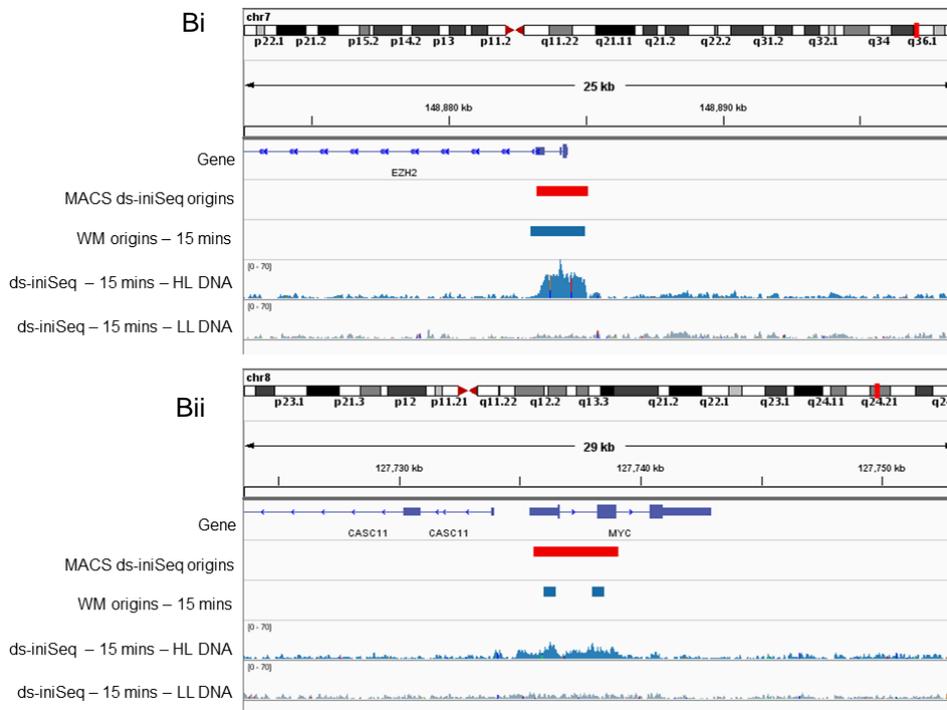
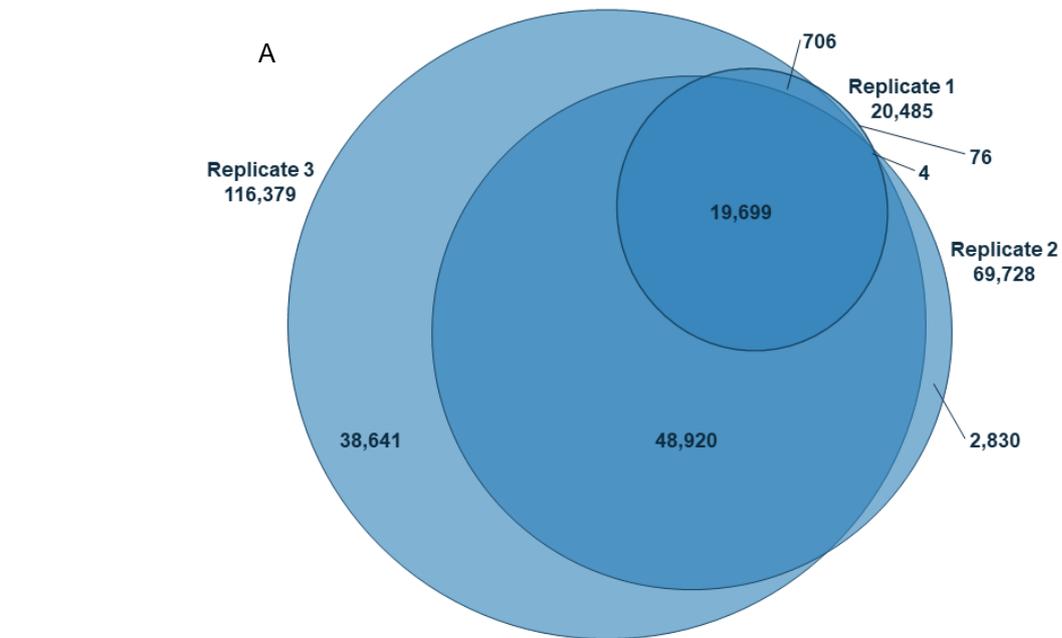
The IGV image of a wider origin (identified by MACS) at the EMILIN3 gene showed similar levels of HL DNA enrichment in both the ds-inoSeq and ds-eloSeq samples (Fig.7.6Bi) but did not appear to be substantially more enriched compared to the neighbouring regions.

The nHL–nLL profile of the ds-inoSeq and ds-eloSeq samples were plotted at this EMILIN3 gene origin +/-100Kb (Fig.7.6Bii), revealing a peak at the origin site and a substantial peak ~50Kb upstream in the ds-inoSeq sample. By contrast no real observable peaks occurred at

the origin site in the ds-eloSeq sample, although a wider but shallower region of nHL-nLL occurred over the origin (~25Kb upstream to ~25Kb downstream). There was also a peak in the ds-eloSeq sample that corresponded to the substantial ds-iniSeq peak ~50Kb upstream of the EMILIN3 origin. The peak ~50Kb upstream of the EMILIN3 origin indicated a neighbouring discrete origin.

### *7.3.2 WM origin calling in 15-minute ds-iniSeq samples*

After establishing that origins could be observed in the nHL-nLL profiles on IGV, we called origins in the ds-iniSeq samples and compared them to the ds-iniSeq origins called by MACS. Using the WM method, we called sites (500bp windows) as potential origins when their nHL-nLL values were greater than 0.00003(AU) (in all three ds-iniSeq replicates). I performed an overlap analysis of the WM origin sites found in the three replicates (Fig.7.7A) and compared them to ds-iniSeq origins called by MACS (Fig.7.7B/C).



**Figure 7.7:** (A) The three-way overlap of the origins called by the WM method in replicates 1, 2 and 3 or the 15-minute ds-iniSeq samples. (B) The IGV images of the sequenced HL and LL DNA (replicate 1) files of replicate 1 of the 15-minute ds-iniSeq samples and the corresponding origin site/sites called by MACS peak caller and the WM method at the EZH2 promoter (i) and the MYC gene (ii). (C) The overlap analysis of the ds-iniSeq origins (replicate 1 origins found in replicates 2 and 3) called by MACS and the replicate 1 origins found in replicates 2 and 3 called by the WM method.

Using the WM method for peak calling in the ds-iniSeq control samples, 20,485, 69,728 and 116,379 origins were identified in replicates 1, 2 and 3 respectively (Fig.7.7A). Comparatively, 15,056, 78,102 and 111,819 origins were identified by MACS in replicates 1, 2 and 3 respectively (Fig.4.15).

Upon overlap analysis, I found 19,699 WM origins from replicate 1 overlapping WM origins present in replicates 2 and 3 (now named “WM origins-15mins”) (Fig.7.7A). There were 14,126 ds-iniSeq origins identified by MACS from replicate 1 overlapping ds-iniSeq origins present in replicates 2 and 3 (now named “MACS ds-iniSeq origins”) (Fig.4.15). Clearly there were more WM origins-15mins than MACS ds-iniSeq origins, demonstrating that the WM method called more potential origin sites than the MACS peaks caller.

I compared the WM origins-15mins with the MACS ds-iniSeq origins, in order to assess the concordance between the two different origin calling methods. An IGV image showed that the WM origins-15mins and the MACS ds-iniSeq origins at the EZH2 promoter were of a similar size and location corresponding to the substantial enrichment of read accumulation in the HL DNA (Fig.7.7Bi). An IGV image showed a wider MACS ds-iniSeq origin when compared to two smaller WM origin-15mins at the MYC gene, where there is an established initiation zone (Fig.7.7Bii). In the grouping distance titration (Chapter 4) I performed on the MACS peak calling, I showed the MYC initiation zone as an example in IGV (Fig.4.12A) which also initially identified two separate potential origin sites prior to grouping, but they were wider than these WM origins-15mins. These differences could indicate that the WM called origins may need grouping.

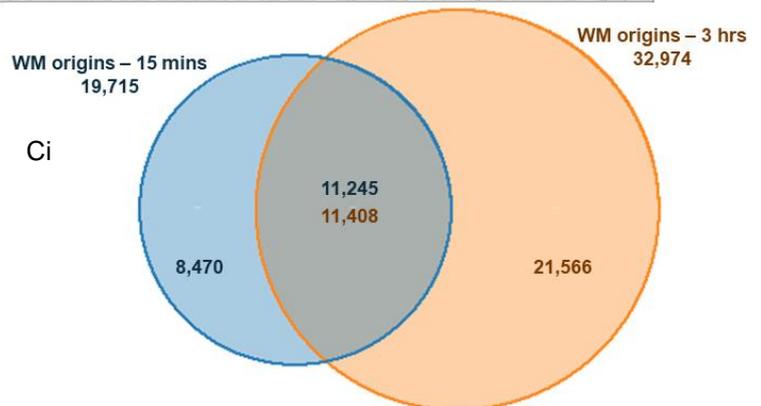
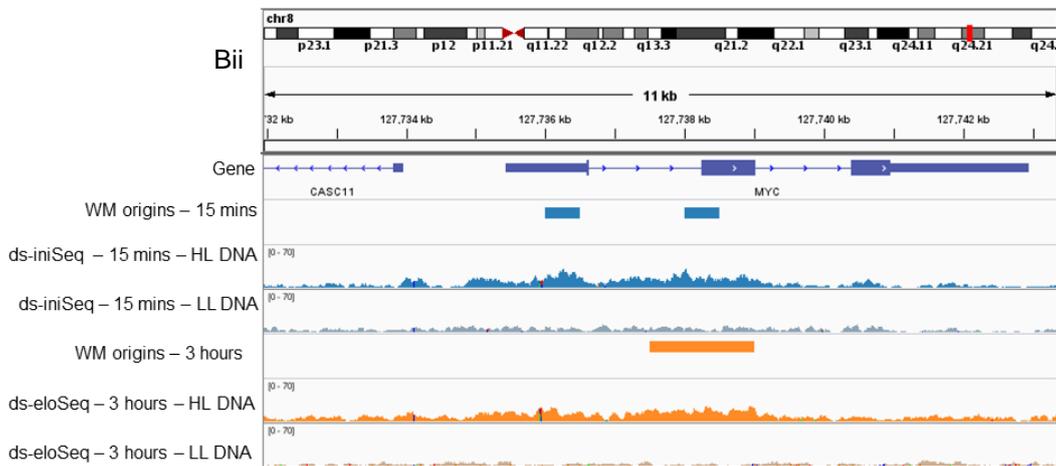
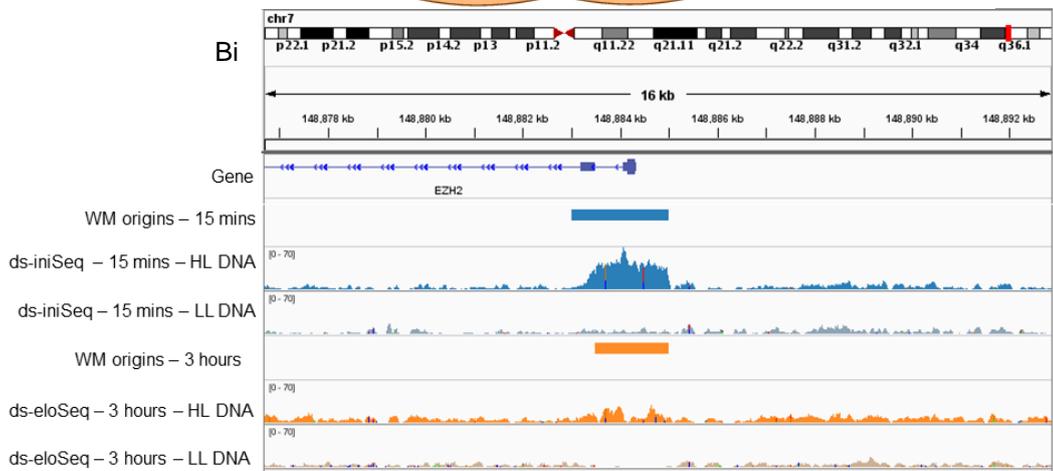
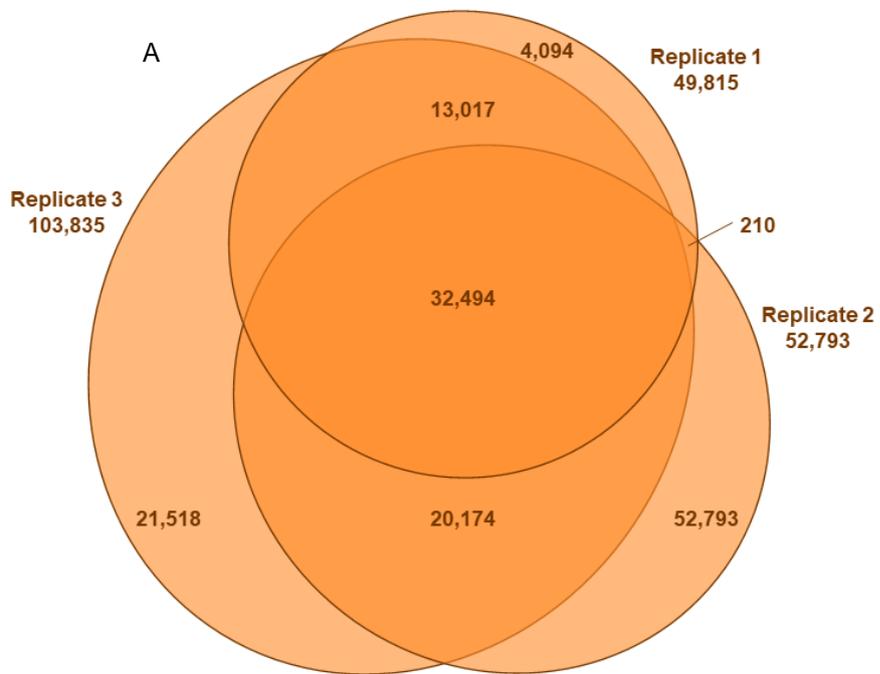
The overlap analysis of the WM origins-15mins and MACS ds-iniSeq origins showed that 73.1% WM origins-15mins overlapped with 89.4% MACS ds-iniSeq origins: thus, showing a very high concordance between them.

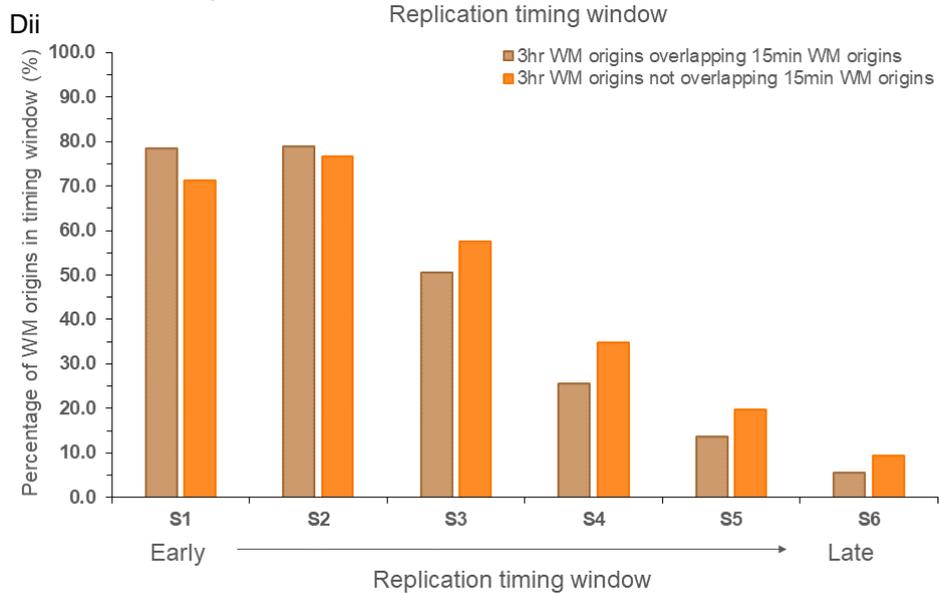
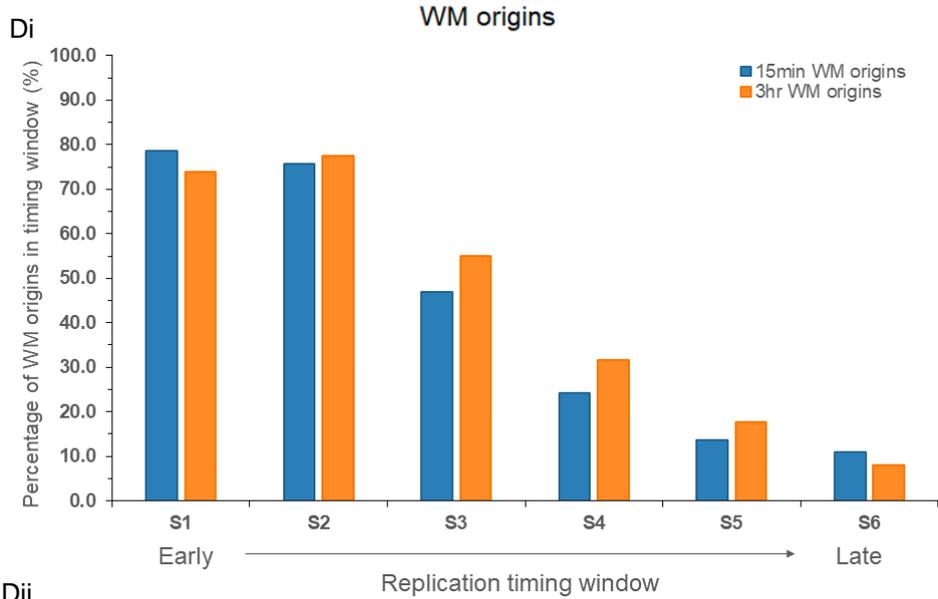
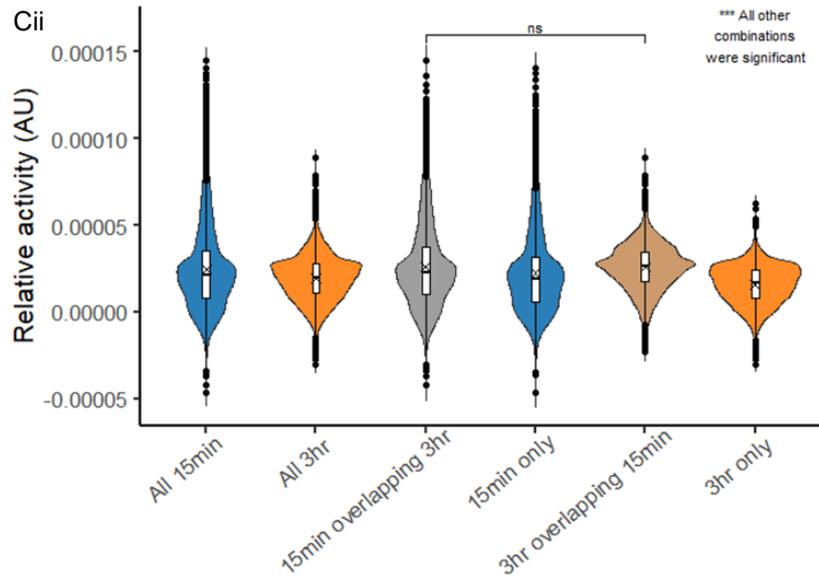
Firstly, these findings made me more confident that the MACS peak caller was mostly identifying the correct origins sites, as both WM and MACS used different methods to call mostly the same sites. Secondly, I concluded that the WM method of origin calling was able to identify origins with sufficient accuracy to use in the ds-eloSeq samples for the identification of origins in the 3-hour samples.

### *7.3.3 WM origin calling in 3-hour ds-eloSeq samples*

We used the WM method of origin peak calling to identify origins in replicates 1, 2 and 3 of the ds-eloSeq control data, on which I performed an overlap analysis to generate a list of potential origin sites that were present in replicate 1 which overlapped sites in replicates 2 and 3 (now named WM origins-3hours) (Fig7.8A). I compared the WM origins-3hours with

the WM origins-15mins to identify those potential origin sites that may have fired after the initial 15-minute replication reaction (Fig.7.8B/Ci). I assessed the quantification of the relative normalised activities of these potential origin sites (Fig.7.8Cii) and compared these WM origins-3hours and WM origins-15mins with replication timing windows to establish if there was an expected shift in origin firing to later in replication (Fig.7.8D).





**Figure 7.8:** (A) The three-way overlap of the origins called by the WM method in replicates 1, 2 and 3 for the 3-hour ds-eloSeq samples. (B) The IGV images of the sequenced HL (orange) and LL DNA (grey orange) files of replicate 1 of the 3-hour ds-eloSeq samples and the sequenced HL (blue) and LL DNA (grey blue) files of replicate 1 of the 15-minute ds-iniSeq samples, and the corresponding sites called by the WM method (orange = 3-hour ds-eloSeq samples; blue = 15-minute ds-iniSeq sample) at the EZH2 promoter (i) and the MYC gene (ii). (Ci) The overlap analysis of the replicate 1 origins overlapping replicates 2 and 3 origins in the 3-hour ds-eloSeq samples called by the WM method (WM origins – 3 hours; orange) with the replicate 1 origins overlapping replicates 2 and 3 origins in the 15-minute ds-iniSeq samples called by the WM method (WM origins – 15 mins). (Cii) The relative activities (normalised read count of the HL-LL, above the threshold for origin calling) of all the WM origins – 3 hours (All 3hr) and WM origins – 15 mins (All 15min) and the activities of those WM origins – 15 mins that did (15min overlapping 3hr) and did not (15min only) overlap the WM origins – 3 hours, and the activities of those WM origins – 3 hours that did (3hr overlapping 15min) and did not (3hr only) overlap the WM origins – 15 mins. An ANOVA and subsequent Tukey's post-hoc test were performed to assess significance; the Tukey's test results are shown on the plot "ns" indicates a not significant result; all other tests were highly significant \*\*\* where  $p < 0.001$ . (Di) The percentages of WM origins – 15 mins and the WM origins – 3 hours that were found in each replication timing window (S1 (early replicating) to S6 (late replicating)) from HeLa cells (8). (Dii) The percentages the WM origins – 3 hours that did and did not overlap the WM origins – 15 mins found in each replication timing window (S1 (early replicating) to S6 (late replicating)) from HeLa cells (8).

I identified 49,815, 52,793 and 103,835 origin sites by WM in replicates 1, 2 and 3 respectively of the ds-eloSeq samples (Fig.7.8A). Of these 32,494 (65.2%) sites in replicate 1 overlapped origins sites in replicates 2 and 3.

The IGV image showed the HL DNA and LL DNA of replicate 1 ds-iniSeq and ds-eloSeq control reactions and the corresponding WM called potential origin sites at the EZH2 promoter origin (Fig.7.8Bi). The EZH2 promoter origin was called in both the 15-minute ds-iniSeq and the 3-hour ds-eloSeq samples. The WM origin-3hours was shorter than the corresponding WM origin-15mins.

The IGV image showed the HL DNA and LL DNA of replicate 1 ds-iniSeq and ds-eloSeq control reactions and the corresponding WM called potential origin sites, at the MYC replication/initiation zone (Fig.7.8Bi). There was an enrichment of HL DNA across the whole MYC initiation zone in both the 15-minute ds-iniSeq and 3-hour ds-eloSeq samples. There were two smaller WM origins-15mins and only the left origin was also called as a WM origin-3hours which was wider than the corresponding WM origin-15min.

My overlap analysis showed that 57.0% of the 19,715 WM origins-15mins overlapped 34.5% of the 32,974 WM origins-3hours (Fig.7.8C).

We were able to quantify the relative activity of the WM origins by determining the normalised read count of the nHL–nLL that were present above the threshold of 0.00003, thus quantifying the significant activity about the background. Fig.7.8Cii showed the activities of all the WM origins-15mins had a larger range and significantly higher mean than all the WM origins-3hours.

I further analysed the difference in activities of the WM origins-15mins that did and did not overlap with the WM origins-3hours, and *vice versa*. I discovered that the relative activities of the WM origins-15mins that overlapped with WM origins-3hours and the relative activities of the WM origins-3hours that overlapped with WM origins-15mins were not statistically different from one another, thus indicating that these common origin sites possessed the same probability of firing in both the 15-minute ds-iniSeq and 3-hour ds-eloSeq samples. The WM origins-15mins unique to the 15-minute ds-iniSeq sample and the WM origins-3hours unique to the 3-hour ds-eloSeq sample were significantly less active than the origins common to both samples, suggesting that they possessed a lower probability of firing. In the WM origins-3hours unique to the 3-hour ds-eloSeq sample some of those origins may include those that fired later in replication compared to the early firing origins found in the 15-minute ds-iniSeq samples.

The expectation is that the origins identified in the 3-hour ds-eloSeq samples would contain origins that fired later in replication than the early firing origins in the ds-iniSeq samples. I compared the WM origins-15mins and the WM origins-3hours with the replication timing windows identified by repli-seq (8). The data in Fig.7.8Di showed a slight shift towards mid-replication timing windows by the WM origins-3hours compared to WM origins-15mins.

The WM origins-3hours that were also found in the 15-minute ds-iniSeq samples (ie overlapping WM origins-15mins), were likely to have fired in the first 15-minutes in both reactions and thus, predominantly, represent early firing replication origins. I compared the WM origins-3hours that did and did not overlap the WM origins-15mins with each replication timing window (Fig.7.8Dii). I found that both groupings had a bias for early-replication timing windows, but the WM origins-3hours unique to the 3-hour ds-eloSeq samples had a greater occupancy at mid- to late- replication timing windows. These data indicated that the WM called origins that fired in the 3-hour ds-eloSeq samples did contain a higher proportion of origins occupying later replication timing windows.

In Fig.7.6-8, I have presented the results of a new method (Wilkes-Mookerjee) for origin calling in these ds-iniSeq and ds-eloSeq samples, with the aim of being able to call activated replication origins in 3-hour ds-eloSeq samples which contain origin firing and replication elongation. WM can identify sharp discrete peaks in the ds-eloSeq samples (containing an extended 3-hour replication reaction), which were indicative of short, sharp replication origin sites that fired, rather than broader regions of replication that would be expected from replication elongation resulting from replication fork progression. The origins called by WM can now be used to identify subsequent elongation sites in the same samples. I was also able to quantify the relative activity of the origins called by WM, and once we have further developed the method to identify elongation sites away from the origin sites, it will be possible to quantify them in the same way; namely from the read count of the nHL-nLL above

the threshold. These could then be used to identify and analyse origins and elongation sites in ds-eloSeq samples from the control, mock-depleted, Y RNA-depleted and xNuRD addition experimental conditions.

## **Chapter 8: A role for the chromatin remodelling complex, Polycomb Repressive Complex 2 (PRC2), in human DNA replication initiation**

### **8.1 Previous work**

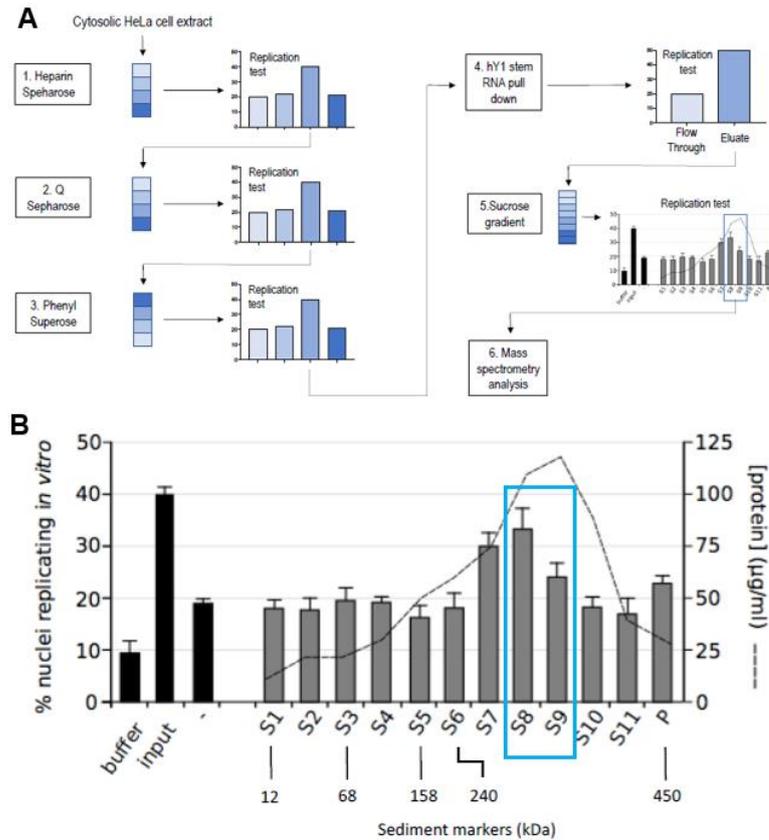
hY RNAs are essential factors in human DNA replication initiation (1). As RNAs in the cell are almost exclusively associated with proteins (2), it is highly likely that these hY RNAs are associated with proteins/protein complexes. The next logical step is to identify Y RNA-binding proteins that are required for DNA replication.

M. Kowalski, (previous PhD student; Krude lab), interrogated Y RNA-binding proteins that were required for DNA replication. She fractionated human cytosolic extract (used in the human cell-free system) through a series of steps (Fig.8.1A) whilst tracking the DNA replication activity; the cytosolic extract was subjected to a fractionation step, and each fraction tested on the human cell-free system. The fractions that resulted in DNA replication activity were then used in the next fractionation step (3).

Integral to this fractionation scheme was a Y RNA pulldown. The upper stem of the hY1 RNA, which is both necessary and sufficient for DNA replication initiation, was bound to agarose beads. The partially fractionated cytosolic extract was incubated with the hY1-beads to separate the upper stem hY1 binding proteins from the other proteins (not associated with hY1). These were tested on the cell-free system, revealing that the fraction containing the hY1 binding proteins possessed replication activity (3).

This active fraction underwent a final fractionation step via a sucrose gradient, which separated the proteins/protein complexes by molecular weight. Two of the fractions (Fig.8.1B) were found to possess a DNA replication activity, and corresponded to molecular weight of 240 – 450 kDa (3).

Proteins from The active fraction S8 and inactive fraction S9 (Fig.8.1B) were identified by mass spectrometric analysis, allowing the inactive fraction to be removed from the active fraction; resulting in a list of proteins that bind to the Y1 RNA upper stem and may be involved in DNA replication initiation (3).



**Figure 8.1:** A schematic diagram of the biochemical fractionation of Y RNA-binding proteins with a replication activity. (A) Cytosolic extract from HeLa cells were subject to the following fractionation steps: heparin sepharose, Q sepharose, Phenyl superose, a hY1 stem RNA pull down and a sucrose density gradient. After each step, the resultant fractions were tested on the human cell-free system. (B) The fractions from the final sucrose density gradient were tested in the human cell-free system and showed a replication activity at fractions 7 and 8 which are consistent with a molecular weight between 240kDa and 450kDa (protein concentration of each fraction was also indicated). The active fraction 8 and the inactive fraction 9 were subjected to mass spectrometric analysis (highlighted in pale blue).

*This figure was extracted from the PhD thesis of M. Kowalski (3) who conducted this work.*

## 8.2 Polycomb-repressive complex 2 – identification and introduction

### 8.2.1 Identification

The mass spec. of the fractionated cytosol generated lists of ~500 potential Y RNA-binding, DNA replication proteins (3). From these lists, I identified the following proteins: EZH1/2, SUZ12, EED, RbAp46/48 (RBBP 7/4) and AEBP2 (Table 1).

Rank in active fraction		Gene Name	Protein name	Mw (kDa)
Replicate 1	Replicate 2			
65	84	EZH2	Enhancer of Zeste Homolog 2	85
140	217	EZH1	Enhancer of Zeste Homolog 1	85
53	81	SUZ12	Suppressor of Zeste 12	83
42	102	EED	Embryonic Ectoderm Development	50
117	243	RbAp46/ RBBP7	Retinoblastoma-Binding Protein 7	47.8
46	60	RbAp48/ RBBP4	Retinoblastoma-Binding Protein 4	47.6
92	141	AEBP2	Adipocyte Enhancer-Binding Protein 2	45

**Table 8.1:** The core subunits and the cofactor of PRC2 complex identified in the mass spectrometric data generated by M. Kowalski. Proteins are listed in order of decreasing molecular weight. Their full name, gene name and rank number in the mass spec. data of the active fraction for 2 replicates are also documented.

The EZH1/2, SUZ12, EED and RbAp46/48 are 4 core protein subunits, and AEBP2 is one common co-factor of the chromatin remodelling complex Polycomb-repressive complex 2 (PRC 2) (4,5).

EZH2, SUZ12, EED and RbAp48 (aka RBBP4) are present in both replicates of the mass spec. data, within the first ~100 ranked proteins (EED was ranked 102 in the 2<sup>nd</sup> replicate). EZH2 and RbAp48 are subunits for which there are alternative isoforms that can form PRC2, in replacement of EZH2 and RbAp48. These isoforms are EZH1 and RbAp46 (aka RBBP7) and were also present in both replicates of the mass spec. data but are not as abundant (higher ranked protein value) as their corresponding isoforms (EZH2 and RbAp48 respectively).

The commonly found cofactor of PRC2, AEBP2 was also present in both replicates, with a ranked abundance that was greater than the EZH2, SUZ12, EED and RbAp48 but lower than the alternative isoforms, EZH1 and RbAp46. Other well documented cofactors, including JARID2, PLC1/2/3, EPOP and PALI1/2 (4,5) were not present in either set of mass spec. data.

It is notable that the total molecular weight of the 4 core subunits is ~310kDa, and ~355kDa with the addition of the cofactor AEBP2. Both molecular weights are consistent with the molecular weights of the active fractions on the sucrose gradients generated by M. Kowalski (240-450 kDa).

The presence of all the core subunits and a cofactor of the PRC2 complex implies that the PRC2 complex is likely to be bound to Y RNAs as a whole complex and is a promising candidate to investigate as a Y RNA-binding DNA replication factor.

### *8.2.2 Introduction – PRC2 background*

PRC2 is a member of the Polycomb group (PcG) proteins, which have two catalytically distinct complex groups (6); Polycomb Repressive Complex 1 (PRC1) compacts chromatin and catalyses histone H2A lysine 119 mono-ubiquitination (H2AK119ub) (7,8); whilst PRC2 catalyses the trimethylation of histone H3 lysine 27 (H3K27me3), and possesses the additional ability to both mono- and di- methylate H3K27 (5,9,10).

These PcG proteins are required for epigenetic regulation through the modification of cellular transcriptional landscapes and their activities are associated with transcriptional silencing often found in heterochromatin (6,11).

PRC1 consists of the RING-type E3 ubiquitin ligase (Ring1A/B and Bmi1 or one of six PCGF subunits), CBX and PHC (canonical PRC1) or RYBP/YAP2 (non-canonical PRC1) (12,13). Interestingly, PRC1 remains associated with chromatin throughout DNA replication (14), however, none of the PRC1 subunits were found in Kowalski's hY RNA-binding mass spec. data (3). The product of PRC1 catalytic activity, H2AK119ub, is known to recruit PRC2. Similarly, the product of PRC2 catalysis, H3K27me3, can be bound by the PRC1 subunit, CBX, demonstrating the cooperation between these two complex classes in the repression of transcription and compaction of chromatin; PRC1 and 2 are often required together for gene repression (6,8).

PRC2 consists of the four core subunits, EZH1/2, SUZ12, EED and RbAp46/48 (aka RBBP 7/4), and a variety of cofactors, including JARID2, AEBP2, PCL 1/2/3 (aka PHF1/MTF2/PHF19), EPOP and PAL1/2 (4,5).

EZH2 is the 751 amino acid, enzymatic subunit of PRC2 (Fig.8.2A). In addition to the highly conserved and methyltransferase domain, Su(var) 3-9 enhancer of zestetrithorax (SET), EZH2 is comprised of five further domains: SANT1- binding domain (SBD); WD40 binding domain, which is the binding domain for the EED subunit (EBD); two SWI3-ADA2-N-CoR-TFIIB [SANT] domains (SANT1 and SANT2); and the cysteine-rich domain (CXC). Further to these key domains, EZH2 also contains a  $\beta$ -addition motif (BAM), a SET activation loop

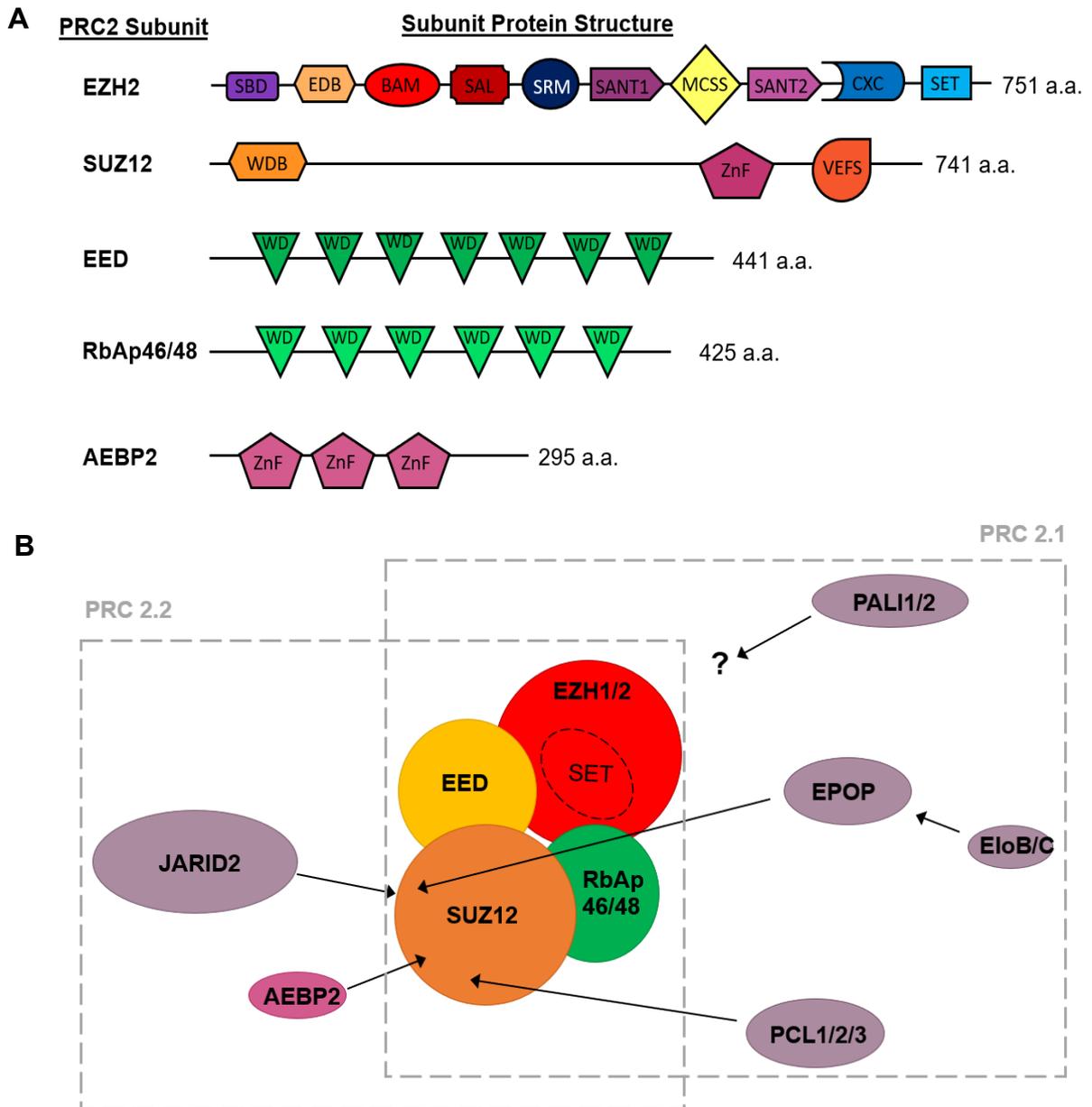
(SAL), a stimulation-responsive motif (SRM), a post-SET domain and the motif connecting SANT1 and SANT2 (MCSS) (5,6,15,16).

EZH1 is a paralog of EZH2 and can substitute for EZH2 in a mutually exclusive manner in the PRC2 complex (17). Crucially, EZH1 contains the catalytically active SET domain (18). When compared to PRC2 containing EZH2, PRC2 containing EZH1 has a lower methyltransferase activity, but a greater affinity for nucleosomes and can compact chromatin independently from its catalytic activity (7,19). Additionally, unlike PRC2-EZH2 the presence of EZH1 in PRC2 results in the inability of this PRC2 to be stimulated/activated by its own product (H3K27me3) (5,20). Studies have suggested that although PRC2 comprised of EZH1 and EZH2 can mono- and di-methylate H3K27, only PRC2-EZH2 can be allosterically modulated to enable it to go on to tri-methylate its substrate. It is believed that this difference in allosteric activation results from differences in the stimulation responsive motif (SRM) in the two subunit isoforms. However, it was found that the binding of the AEBP2 cofactor to EZH1 containing PRC2 can stimulate its activity to levels similar to that of AEBP2 stimulation of EZH2 containing PRC2 (17). Moreover, PRC-EZH2 is predominantly expressed in proliferating cells, whereas PRC2-EZH1 is highly expressed in non-dividing cells (10).

Finally, PRC2 complexes can form homo- and hetero-dimers of the EZH1 and EZH2 forms. It has been suggested that this dimer formation and the ratio of EZH1 and EZH2 may modulate histone methyltransferase activity in a cell-dependent manner (20).

Although EZH2 (and EZH1) possesses the catalytic activity of this protein complex, for it to efficiently carry out its histone methyltransferase activity, it requires the binding of two other core subunits, SUZ12 and EED (20–23). In fact, when in isolation EZH2 will exert autoinhibition, where its conformation allows the post-SET domain to interact with the lysine binding region of the active site, thus blocking its activity (20,24–26).

EED is the subunit responsible for stimulating EZH2 histone methyltransferase activity (27) and is involved in the recruitment of PRC2 to H3K27me3 sites (28). Studies have shown that the interaction of EED with the PRC2 product, H3K27me3, results in the stimulation of the PRC2 complex's catalytic activity (27). As a result, EED constitutes an essential component of the PRC2 complex (5,6).



**Figure 8.2:** (A) The functional domains, motifs, and amino acid (a.a.) length of the PRC2 subunits. Domains and motifs comprise: SANT1-binding domain (SBD), WD40/EED binding domain (EBD),  $\beta$ -addition motif (BAM), SET activation loop (SAL), stimulation responsive motif (SRM), SWI3-ADA2-N-CoR-TFIIIB domains 1 and 2 (SANT1 & SANT2), motif connecting SANT1 and SANT2 (MCSS), cysteine-rich domain (CXC), Su(var) 3-9 enhancer of zestetrithorax (SET), WD40 binding domain (WDB), Zinc-finger region (ZnF), VRN2-EMF2-FIS2-SU(Z)12 (VEFS) and WD40 repeat domains (WD). These core subunits come together to form the PRC2 complex (B) where two subcomplexes are formed with the association of PRC2 and some mutually exclusive cofactors (the catalytically active SET domain is highlighted). (I generated these diagrams based on the information in (4-6)).

EED is a 441 amino acid protein, consisting of seven WD-40 repeat motifs (WD), which results in a seven-bladed  $\beta$ -propeller structure (Fig.8.2A) (29). These are 4 isoforms of mammalian EED, all of which have been found to perform the same role in the PRC2 complex *in vivo* and *in vitro* (30).

SUZ12 binds both EZH2 and EED (6), is the subunit essential for maintaining the structural integrity of PRC2 (Fig.8.2A) (23) and is involved in chromatin binding (31) through its interaction with cofactors (32). It is a 741 amino acid protein that contains a WD-40 domain (WDB), a zinc-finger region and most importantly, a conserved VRN2-EMF2-FIS2-SU(Z)12 (VEFS) domain (5,6). This VEFS domain is responsible for stably binding EZH2 and is involved in facilitating the allosteric stimulation of PRC2's histone methyltransferase activity (33). The interaction of this VEFS domain with EZH2 and EED has been designated as the PRC2 minimal core as it is sufficient to carry out PRC2's methyltransferase activity (5).

SUZ12 also binds to RbAp46/48 (34,35) and can mediate the interaction of PRC2 with its cofactors, including JARID, AEBP2, PLC1/2/3 and EPOP (Chammas *et al*, 2019). In the absence of SUZ12, EZH2 continues to bind to EED but they are unable to bind to other proteins/cofactors (36). In addition, it has a high affinity for non-coding RNAs (37) and appears to be sufficient for RNA binding (38).

Finally, RbAp46/48 (RBBP7/4) are roughly 425 amino acids and comprise 6 WD-40 repeat domains (WD), which ultimately fold into a 7-bladed  $\beta$ -propeller structure (Fig.8.2A). Although the RbAp46/48 subunits are not essential for PRC2 activity, they have been found to interact with histones, in particular with the H3-H4 heterodimer (6,39). As a result, they are believed to be involved in the recruitment of PRC2 to nucleosomes (6,40).

These PRC2 subunits have been found to play roles in structural support and protein-protein interactions in other chromatin remodelling complexes (41–46). RbAp46/48 have been found in the Y RNA independent DNA replication factor, Nucleosome remodelling and histone deacetylase complex in early *X. laevis* embryos (xNuRD) (47).

Of the core complex, there are four important structural lobes (that relate to function), Middle, Regulatory, Docking and Catalytic lobes (5).

The catalytic, regulatory and middle lobes work together to bring about the catalytic activity of PRC2 (5). The catalytic lobe consists of the CXC and SET domains of EZH2, which form the active site, comprising two pockets; the first is a highly hydrophobic channel that interacts with the aliphatic chain of the lysine substrate, and the second is situated at the end of the hydrophobic channel, accommodating the methyl group donor, S-adenosyl methionine (SAM) (24–26).

The regulatory lobe is in close contact with the catalytic lobe and consists of the EBD and BAM regions of EZH2 and part of the EED  $\beta$ -propeller (29,48). EED is the subunit that binds tri-methyl peptides, resulting in a conformational change of the PRC2 complex and increased efficiency of catalysis over and above PRC2's basal activity (20,21,48,49). The middle lobe

consists of part of EZH2 and the region of SUZ12's VEFS domain that is between the regulatory and catalytic lobes, and stabilises the active site of EZH2 (5,48,50).

The docking lobe, considered to be responsible for recruitment of proteins to PRC2, comprises the N-terminal of SUZ12 (35,51). This region acts as a docking platform for additional factors, including RbAp46/48 to associate with (35). Interestingly, the binding of SUZ12 with RbAp46/48 results in the inhibition of RbAp46/48 binding with its known nuclear factors, including nucleosomes, which does seem counterintuitive; the reasons for this remain unelucidated (5,31).

PRC2 has been implicated in disease including cancer (6). Genome-wide analyses have shown that PRC2 activity and H3K27me3 is commonly associated with tumorigenesis (52–55). EZH2 is overexpressed in numerous cancers including bladder, breast, colon, lung, pancreatic, prostate and renal cancers, and some sarcomas and lymphomas (56–65). This overexpression is often associated with poor prognosis but presents a potential avenue for therapeutic treatments (6,15,53,55,56,66). PRC2 also appears to be involved in additional cellular function, including development, lymphocyte activation, protecting neurons from neurodegeneration and T-cell stimulation response against tumours (67–70). In addition, high PRC2 expression correlates with cancer cell proliferation (71,72), which may imply a role in DNA replication.

The PRC2 core complex remains consistent, but PRC2 complexes can be subdivided based on mutually exclusive cofactors/facultative subunits, into two functionally different subclasses; PRC 2.1 and 2.2 (Fig.8.2B) (4,5). The presence and location of the subcomplexes is considered to be dependent on cell-type. Subclass PRC2.1 can include PALI1, EPOP, PCLs 1/2/3 (aka PHF1/MTF2/PHF19), whereas subclass PRC2.2 includes JARID1 and AEBP2 (4,5). Notably AEBP2 is the only cofactor found in Kowalski's mass spec. data of Y RNA-binding proteins in a fraction with replication activity.

The cofactors which associate with the PRC2 complex are believed to play a role in regulating the complex's recruitment and activity level (11,73). PRC2.1 can only accommodate one PCL isoform at a time (4). PCL2 has been shown to facilitate PRC2 recruitment to unmethylated CpG islands (CGIs), a genomic feature that PRC2 is known to prefer (74–76). PCL-1 and -3 are expressed in lower levels than PCL2 and can modulate recruitment and activity of PRC2 (4,77,78); PCL1 can bind DNA directly, which increases the time PRC2 occupies the chromatin, thus stimulating H3K27 trimethylation. EPOP and PALI1/2 can bind PRC2 in addition to the PCLs (79–81). EPOP acts as a bridge between PRC2 and EloB/C, which in turn can modulate RNA polymerase II, whereas PALI1 can stimulate EZH2 activity (4,79,81,82).

Finally, the PRC2.2 cofactors JARID2 and AEBP act to cooperatively stimulate EZH2 catalytic activity above its basal rate (4,7,83). JARID2 can be methylated by PRC2 resulting in JARID2K116me3. Although JARID2 binds to SUZ12, the methylated lysine interacts with EED, resulting in a conformational change of PRC2 and ultimately stimulates the catalytic activity of EZH2. This methylated JARID2-EED interaction mimics the interaction of H3K27me3 with EED and is believed to be a potential mechanism by which novel transcriptionally repressive marks (H3K27me3) can be established (5,11,49,83). AEBP2 also stimulates PRC2 catalytic activity via the stabilisation of EZH2 and/or increasing the binding of PRC2 with nucleosomes (22,27,35,83). This increased preference for nucleosome binding is further enhanced if the nucleosomes contain H2AK119ub (deposited by PRC1) (4,84). It has been suggested that only PRC2.2 can bind nucleosomes *in vitro* (5,31).

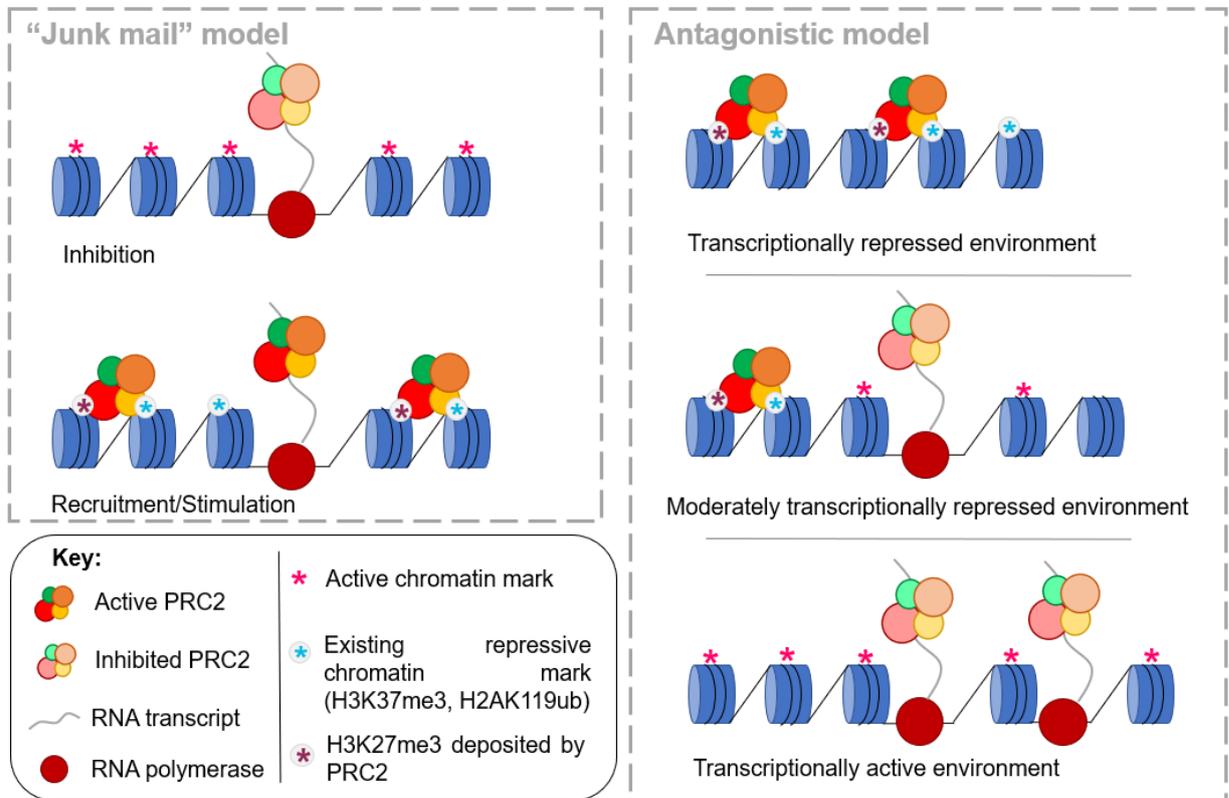
Additional factors that can regulate PRC2 are histone post-translational modifications and RNA (5). H2AK119ub recruits PRC2 to chromatin (6,8) and H3K27me3 stimulates PRC2 activity, which is believed to constitute as positive feedback in order to ensure the maintenance of a repressive chromatin state (27). In addition to these, other histone modifications that mark chromatin repression, including H1K26me3, H3K9me3 and H4K20me3, can be recognised by PRC2 (21,85,86).

The histone modifications H3K4me3 (associated with DNA replication origins) and H3K36me2/3 (associated with DNA damage regulation), which are associated with transcriptionally active chromatin, strongly inhibit PRC2 activity; suggesting an additional mechanism by which transcriptionally active states are maintained (34,86,87). Interestingly, unmodified H3K36 has been found to stimulate PRC2 activity (88).

Despite the lack of a defined RNA binding site (89), PRC2 has been found to associate both promiscuously and more specifically, with in excess of 1,000 RNAs (up to 20% of human long non-coding RNAs (lncRNAs)) (90–93). EZH2 (whose RNA affinity reduces when in complex with PRC2), SUZ12, AEBP2 and JARID2 all have differing binding affinities for RNAs. EED has a low RNA binding affinity and has been found to regulate EZH2-RNA binding affinity (94,95).

It is believed that RNAs play a role in the recruitment and activity of PRC2 (96). In particular, the binding of RNA to PRC2 has been found act as a non-active site inhibitor of PRC2's methyltransferase activity (95,97). lncRNAs have been found to recruit PRC2 to chromatin (98,99). However, evidence has suggested that PRC2 binding of chromatin and RNA is mutually exclusive, thus implying that RNA binding can inhibit PRC-chromatin recruitment (38). Additionally, increasing RNA concentrations can reduce PRC2 association with nucleosomes (97).

Currently, there are two proposed models by which RNAs can regulate PRC2 activity: ‘Junk mail’ and Antagonistic models (Fig.8.3) (100). In the “junk mail” model, PRC2 is inhibited by RNA but this inhibition is overcome in the presence of PRC2 stimulating/ transcriptionally repressive marks, which in turn enables the generation of H3K27me3 (94,96).



**Figure 8.3:** The PRC2 complex is thought to be regulated by RNA transcripts and the currently proposed models are the “junk mail” and antagonistic models. In the “junk mail” model, PRC2 is inhibited by RNA but this inhibition is overcome in the presence of PRC2 stimulating/ transcriptionally repressive marks. The antagonistic model suggests that PRC2 activity is dependent on the transcriptional environment; transcriptionally repressed environments result in the propagation and maintenance of H3K27me3, and in a moderately transcriptionally repressed environment the RNA and repressive chromatin marks compete for PRC2 binding, and in a transcriptionally active environment, PRC2 is inhibited. (I generated these diagrams based on the information in (100)).

The Antagonistic model demonstrates an antagonistic relationship between chromatin and RNA for binding with PRC2. In a transcriptionally repressed scenario, PRC2 binds with chromatin and propagates H3K27me3. In a moderately transcriptionally repressed state, PRC2 is competed for by RNA (which would inhibit PRC2 activity) and repressive marks on chromatin (which would stimulate PRC2 activity) (21,84,100–102). In a transcriptionally active environment RNA is abundant and binds to and inhibits PRC2 (100).

PRC2 has been found to most commonly bind to lncRNAs, including *Xist* and *RepA*, *Tisx* and *HOTAIR* (100). PRC2 binds with the *Xist* and *RepA* RNAs as they are being transcribed and, as a result, directs PRC2 to chromatin where it can identify repressive markers and tri-

methylation of H3K27 is stimulated. PRC2-*Tisx* binding is understood to inhibit PRC2 activity, thus ensuring the transcriptional activity of the active X-chromosome (92,100,103,104). The binding of PRC2 with *HOTAIR* is a suggested example of trans-regulation, where the *HOTAIR* transcript binds with PRC2 which lays down the repressive H3K27me3 mark at a different genomic location (in this case, at the *HOXD* locus) to the *HOTAIR* gene (91,105). PRC2 has been found to localise at both transcriptionally active and inactive sites, however it is catalytically inactive at transcriptionally active chromatin (94).

It was believed that a two-hairpin motif on the RNAs is responsible for their binding to PRC2, however this motif appears to be insufficient for binding (81,93). Furthermore, PRC2 does show a high affinity for G-rich regions in ssRNA and G-quadruplexes (97,106) and PRC2 affinity for RNAs increases as the length of RNAs increase (96). In addition, the affinity of PRC2 for RNAs reduces when it is perfectly double-stranded; a perfect dsRNA rarely occurs *in vivo* (97,106). The introduction of a few helical defects into double-stranded RNA, including small internal loops and bulges, has been found to increase the affinity of PRC2 for double-stranded RNAs (97).

It has been suggested that there is potential for additional unknown RNA binding proteins that associate with PRC2 and mediate its interaction with RNAs (100). RNA binding protein fox 1 homologue 2 (RBFox2) is thought to fulfil this role which facilitates the interaction of *Rep A* and PRC2 (107,108).

Taken together, this information about the PRC2 complex, in particular its role in remodelling chromatin and its ncRNA binding affinities and the mass spec lists, indicate that PRC2 is a viable candidate for a Y RNA-binding, DNA replication initiation factor.

## **8.3 Results**

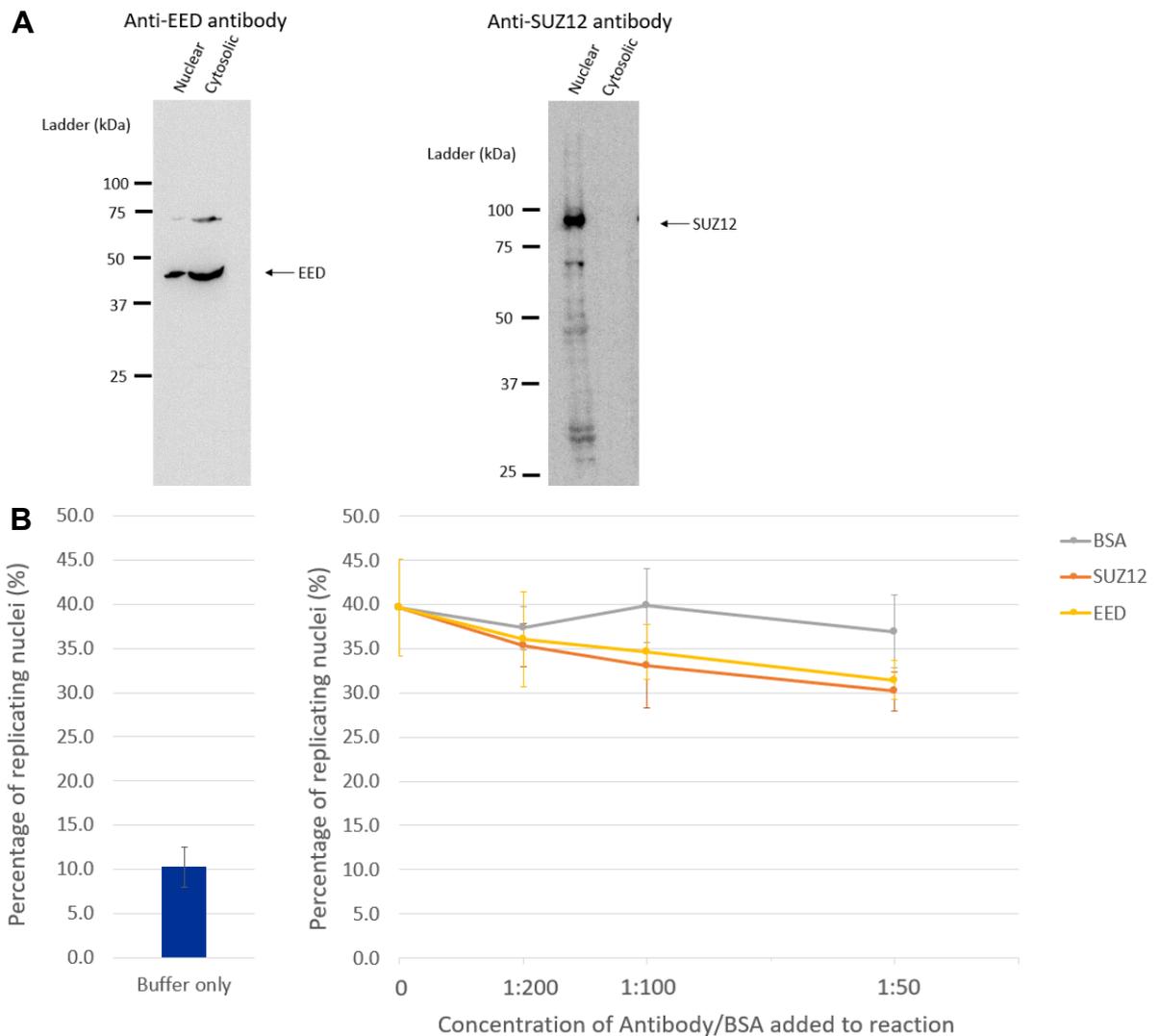
### *8.3.1 This Project*

This work explored the role that the chromatin remodelling complex, PRC2, may play in DNA replication initiation. In order to do this, I performed a number of inhibition studies using the human cell-free system.

I inhibited PRC2 through the treatment of the human cytosolic extract (used in the cell-free system) with antibodies, a chemical inhibitor and through immunodepletion. The resultant extracts were then tested in the human cell-free system. Should PRC2 play a role in DNA replication, the percentage of replicating nuclei should be reduced upon its removal, when compared to the appropriate controls.

### *8.3.2 Antibody treatment of the cytosol*

I assessed the effect of the inhibition of the core structural subunits, SUZ12 and EED by specific antibodies, on DNA replication. I tested the quality of available antibodies and determined the location of these subunits in the cell, through testing nuclear and cytosolic extracts from HeLa cells on a 10% SDS polyacrylamide gel, which was immunoblotted with the anti-EED and -SUZ12 antibodies (Fig.8.4A). I analysed the effect of the inhibition of the PRC2 complex with anti- SUZ12 and EED antibodies in the human cell-free system (Fig.8.4B). I incubated the cytosol with either anti-SUZ12 or anti-EED antibodies or a BSA control (to account for additional protein in the cell-free system), at increasing concentrations. I added this treated cytosol to the cell-free system, where the replication reaction took place, and determined the percentages of replicating nuclei.



**Figure 8.4:** (A) Cytosolic and nuclear extract of HeLa cells were loaded onto a 10% SDS polyacrylamide gel, which then underwent a Western blot, developed using antibodies for the PRC2 subunits, SUZ12 (right) and EED (left), to assess their quality and specificity and to observe the distribution of the subunits within the cell. Bands were present at the documented molecular weight for both EED (49kDa) and SUZ12 (83kDa). (B) Synchronised late G1 template nuclei were incubated with antibody/BSA-treated cytosolic HeLa extract at increasing concentrations of BSA (control) or anti-EED or anti-SUZ12 antibodies (1:200, 1:100, 1:50) and a mixture of dNTPs/NTPs (including a dig-dUTP) in physiological buffer. A negative control where the template nuclei were incubated without cytosol was also conducted (“buffer only”). The nuclei were then spun onto a glass coverslip and stained for DNA and the incorporated dig-dUTP. The percentages of replicating nuclei were found through immunofluorescence microscopy and the mean values, with standard deviations are shown here. A 2-tailed student’s T-test (unequal variance) was performed and demonstrated no significant decrease in replication activity when compared to the BSA controls at corresponding concentrations (n= 3-4).

The Western blot for anti-EED (Fig.8.4A-left) showed two bands at a little under 50kDa, present in both cytosolic and nuclear extracts, which is consistent with the size of the EED protein subunit. They appeared to be in similar intensity in both the nuclear and cytosolic extracts.

The manufacturers of this antibody (Abcam) stated that the antibody detects a band only at 80kDa, despite the molecular weight of EED being 50kDa. This Western blot also showed a very faint and small band in the nuclear and cytosolic extracts respectively at just under 75kDa, which was inconsistent with the observations suggested by Abcam. This secondary band is therefore probably due to non-specific binding rather than the EED protein.

The Western blot for the anti-SUZ12 antibody (Fig.8.4A-right) showed one band at ~80kDa present only in the nuclear extract. This band is consistent with the size of SUZ12 of 83kDa. I therefore conclude that SUZ12 is abundant in the nucleus of the HeLa cells but not in the cytosol.

As SUZ12 is only present in the nucleus but EED is present in both the nucleus and cytosol, it is possible that EED forms part of another protein complex that is present in the cytosol but SUZ12 does not. Alternatively, EED may be present in the cytosol separate from any complex.

Ultimately, these Western blots have demonstrated that the anti -EED and -SUZ12 antibodies are specific for their appropriate proteins and are suitable for use for further analysis in the cell-free system. However, I considered the anti-SUZ12 antibody more reliable than the anti-EED antibody, due to the divergence of my Western blot result from the manufacturer's product description.

Analysis of these antibodies in the cell-free system (Fig.8.4B) showed approximately 10% contamination from S-phase nuclei in the buffer only control. With the addition of untreated cytosolic extract, the percentages of replicating nuclei increased to approximately 40%, indicating that ~30% of late G1-phase nuclei were replicating.

The addition of BSA to the cytosol in these replication reactions, at increasing concentrations (1:200, 1:100, 1:50 dilutions of antibody stock) had a negligible effect on replication and percentages of replicating nuclei were not significantly different when compared to the cytosol only control.

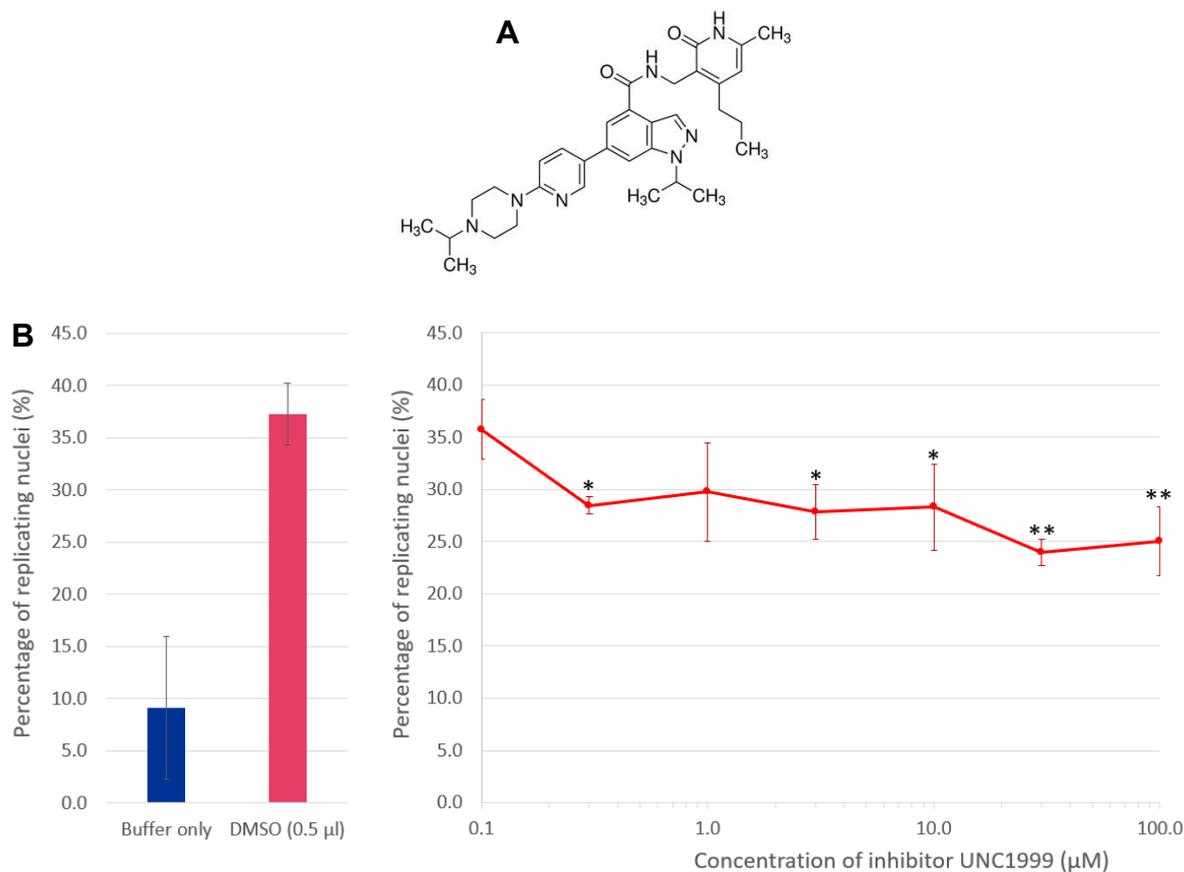
The addition of anti -EED and -SUZ12 antibodies independently to the cytosol in these replication reactions, at increasing concentrations (1:200, 1:100, 1:50 dilutions of antibody stock) showed a very similar pattern. At 1:200, there was an almost negligible reduction in DNA replication activity compared to the equivalent BSA control. At 1:100 and 1:50, there was a small reduction in DNA replication compared to the equivalent BSA controls. There was an observable reduction in DNA replication activity for both the anti-EED and anti-SUZ12 conditions which was not significant compared to the corresponding BSA control (for 1:50,  $p = 0.13$  and  $p = 0.087$  for EED and SUZ12 respectively).

The data presented here, demonstrate that treatment with EED and SUZ12 specific antibodies (at higher concentrations) results in a small but insignificant inhibition of DNA replication in the cell-free system. Although this reduction of the percentages of replicating nuclei is not significant, it indicates that it is worth carrying out further independent methods of inhibition.

### *8.3.3 Chemical inhibitor treatment of the cytosol*

I carried out a chemical inhibition of PRC2 by titrating the small molecular competitive inhibitor of EZH1 and EZH2, UNC1999 (Fig.8.5) at increasing concentrations into the cell-free system (Fig.8.5B). UNC1999 (dissolved in DMSO) inhibits the action of the enzymatic subunits, EZH1 and EZH2, by occupying the protein's binding site for the reaction substrate, SAM (109,110) (Fig.8.5A). I introduced DMSO to an additional replication reaction (at the same volume as UNC1999), to act as a positive control.

The human cytosol was incubated with DMSO or increasing concentrations of UNC1999, for 30mins prior to inclusion in the human cell-free system.



**Figure 8.5:** (A) The molecular structure of the small chemical inhibitor of the PRC2 catalytic subunits, EZH1 and EZH2, known as UNC1999 (structure from manufacturer Sigma-Aldrich (product code - SML0778)). (B) Synchronised late G1 template nuclei were incubated with UNC1999-treated or DMSO-treated (at the same volume that was used for the chemical inhibitor) cytosolic HeLa extract at increasing concentrations (0.1, 0.3, 1, 3, 10, 30 and 100  $\mu\text{M}$ ) and a mixture of dNTPs/NTPs (including a dig-dUTP) in physiological buffer. A negative control where the template nuclei were incubated without cytosol was also conducted ("buffer only"). The nuclei were then spun onto a glass coverslip and stained for DNA and the incorporated dig-dUTP. The percentages of replicating nuclei were found through immunofluorescence microscopy and the mean values, with standard deviations are shown here. A 2-tailed student's T-test (unequal variance) was performed and demonstrated a significant reduction in replication activity of inhibitor concentrations of 0.3, 3, 10, 30 and 100  $\mu\text{M}$  when compared to the DMSO control (n= 3). (\*  $p < 0.05$ ; \*\*  $p < 0.01$ )

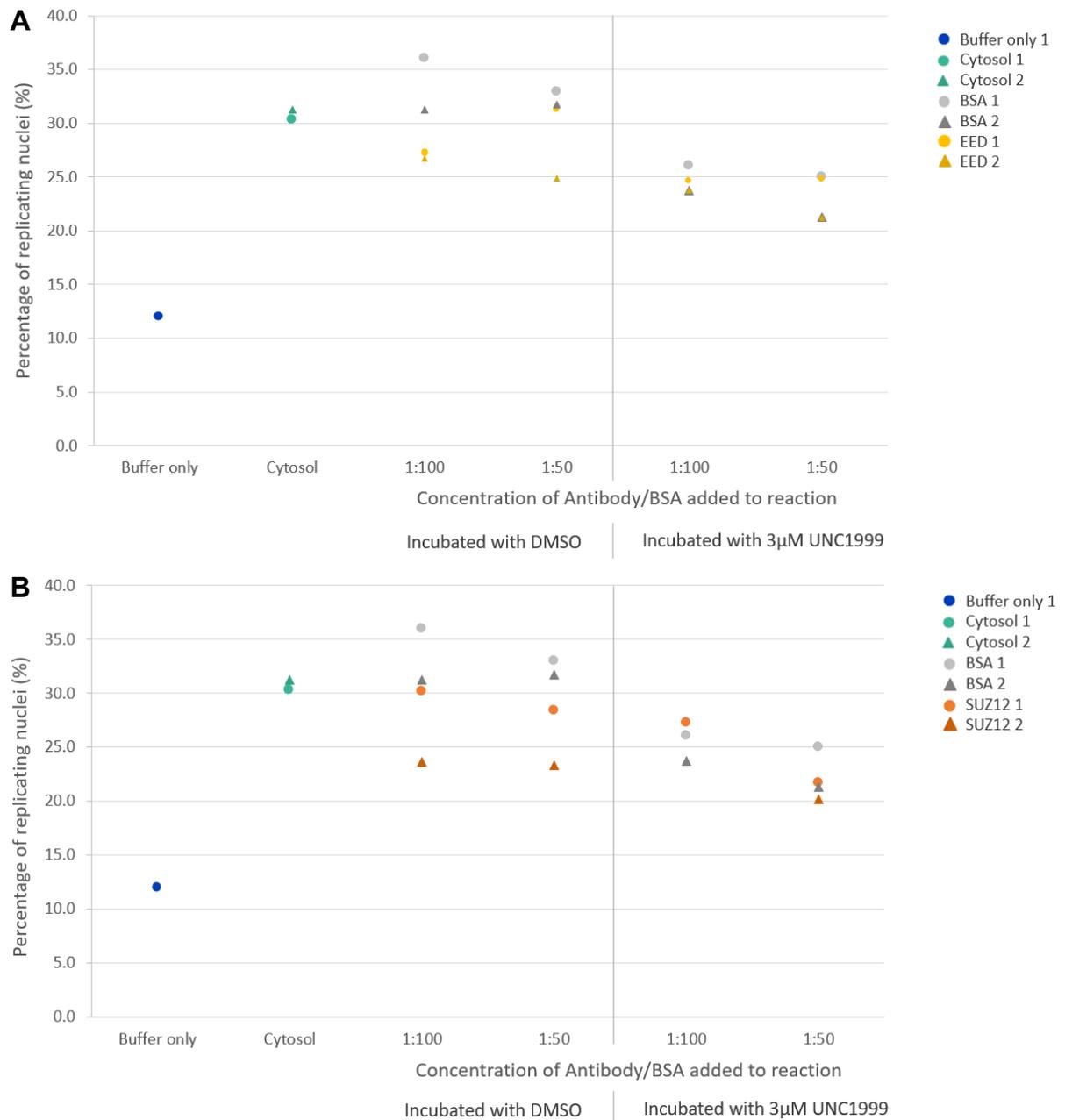
The control reaction in buffer only confirmed the presence of approximately 10% contamination of S-phase nuclei in the template nuclei (Fig.8.5B left). Upon the addition of cytosol pre-treated with DMSO, the percentage of replicating nuclei was increased to approximately 37%. The introduction of 0.1  $\mu\text{M}$  UNC1999 to the cell-free system did not affect replication activity. The introduction of 0.3, 1, 3, 10, 30 and 100  $\mu\text{M}$  of UNC1999 to the cell-free system, resulted in the reduction of the percentage of replicating nuclei by up to 10% (Fig.8.5B right). Of these concentrations, 0.3, 3 and 10  $\mu\text{M}$  UNC1999 showed a significant reduction in DNA replication at the significance threshold of  $p < 0.05$  compared to the DMSO control. Concentrations of 30 and 100  $\mu\text{M}$  UNC1999 showed significant reductions in DNA replication activity at the significance threshold of  $p < 0.01$ , compared to the DMSO control.

Despite these significant results, the inhibition of PRC2 does not reduce DNA replication activity to baseline levels (ie that of the buffer only condition).

These results indicate that the inhibition of the enzymatic subunits of PRC2, EZH1/2 partially inhibits DNA replication *in vitro*.

#### *8.3.4 Antibody and Chemical inhibitor treatment of the cytosol*

As seen above (Fig.8.5), the inhibition of EZH1/2 significantly reduced DNA replication to a partial degree. In order to assess whether the targeting of SUZ12 or EED by specific antibodies in addition to the inhibition of EZH1/2 resulted in a further reduction in DNA replication, I incubated human cytosol with both UNC1999 and anti -EED or -SUZ12 antibodies. DMSO and BSA were used as positive controls in the cell-free system (Fig.8.6).



**Figure 8.6:** Synchronised late G1 template nuclei were incubated with UNC1999-treated (3µM) /DMSO-treated (at the same volume that was used for the chemical inhibitor) and antibody/BSA-treated (concentrations of 1:100 and 1:50) cytosolic HeLa extract and a mixture of dNTPs/NTPs (including a dig-dUTP) in physiological buffer. The antibodies used were for EED (A) and SUZ12 (B). A negative control where the template nuclei were incubated without cytosol (“buffer only”) and a positive control where template nuclei were incubated with untreated cytosol (“cytosol”) were also conducted. The nuclei were then spun onto a glass coverslip and stained for DNA and the incorporated dig-dUTP. The percentages of replicating nuclei were found through immunofluorescence microscopy and the mean values, with standard deviations are shown here. No statistical analysis was performed as n=1-2, however the data tends to indicate that there was unlikely to be an additional cumulative effect of inhibiting more than 1 subunit of PRC2. (n= 1-2)

The combined inhibition of EZH1/2 and EED (Fig.8.6A) indicated ~10% contamination of S-phase nuclei (buffer only) and the increase of the percentage of replicating nuclei to approximately 30% with the addition of untreated cytosol.

The addition of DMSO and BSA resulted in no real difference and the addition of the anti-EED antibody showed a small reduction in the percentage of replicating nuclei, compared to its DMSO/BSA counterpart. The addition of 3 $\mu$ M UNC1999 and BSA resulted in a reduction in percentage of replicating nuclei, which was in line with the reduction observed in the chemical inhibition study. Upon the addition of anti-EED antibody to the UNC1999, no further reduction was observed, even at the highest concentration of antibody.

The combined inhibition of EZH1/2 and SUZ12 presented the same buffer only, cytosol, DMSO/BSA and UNC1999/BSA control conditions (Fig.8.6B), as those in the EZH1/2 and EED study (Fig.8.6A) described above. The addition of anti-SUZ12 antibody to DMSO resulted in a reduction of ~5% of replication, which is in line with the previous antibody inhibition study. As with the addition of anti-EED, the introduction of anti-SUZ12 antibody (at both concentrations) to the UNC1999 in the replication reaction, resulted in no further reduction in replicating nuclei levels over and above that of UNC1999 alone. The inhibition of SUZ12 alone resulted in a reduction of replication activity that was roughly equivalent to that of UNC1999/BSA and UNC1999/anti-SUZ12.

Due to resource constraints, I was only able to carry out two repeats of this experiment. As a result, I was unable to perform a statistical analysis. However, these results strongly suggest that there was no additional effect on DNA replication *in vitro*, when EZH1/2 and SUZ12 or EED were inhibited together.

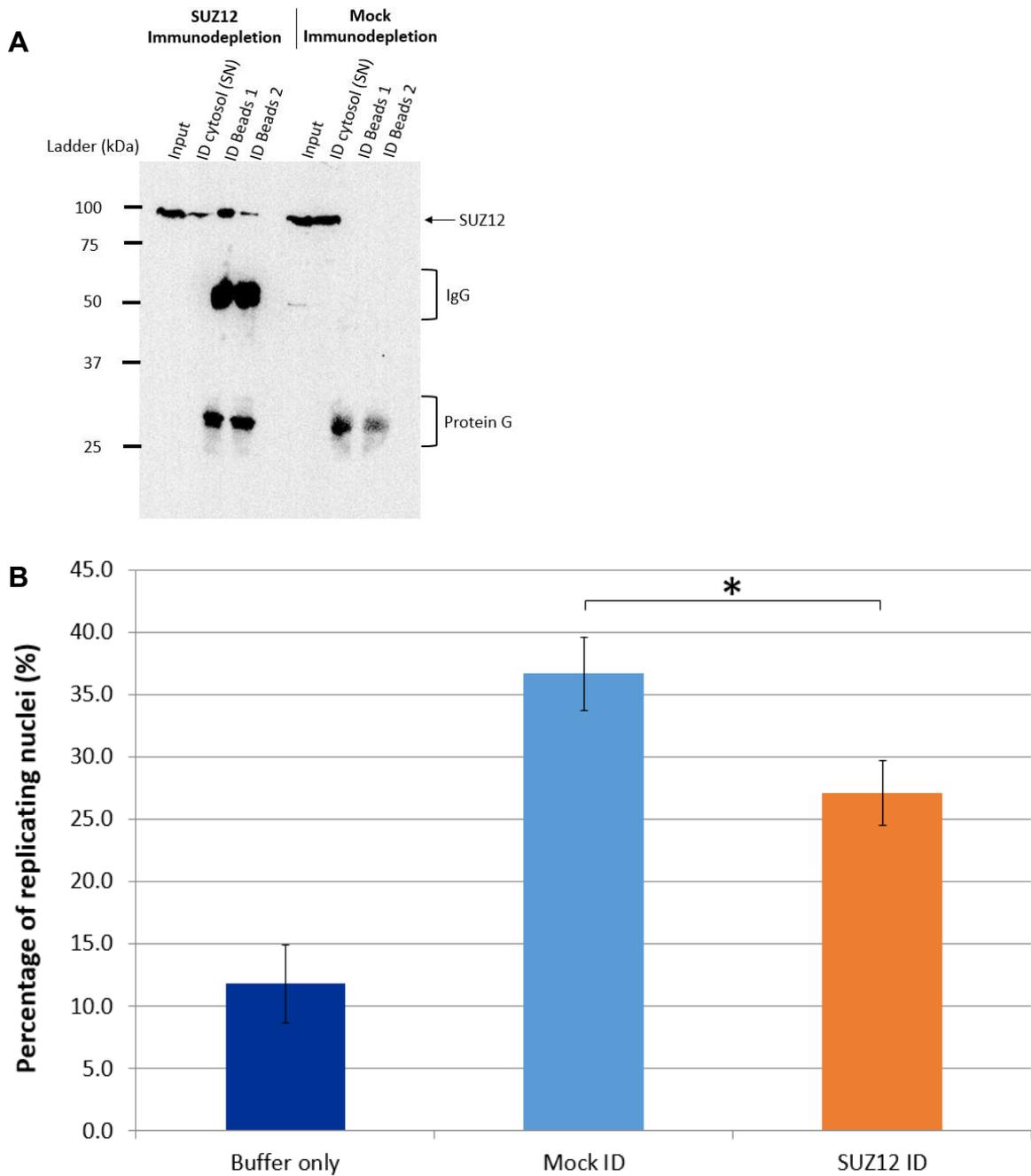
### 8.3.5 Immunodepletion of SUZ12 from the cytosol

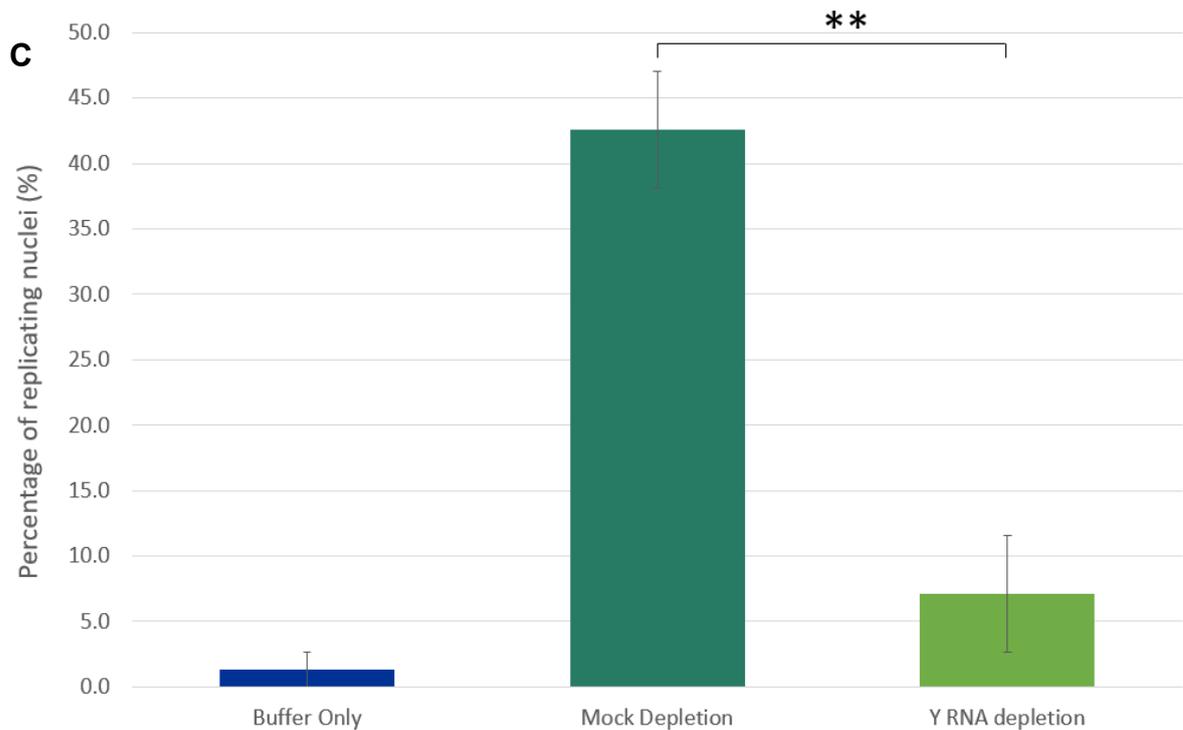
I performed an immunodepletion of the human cytosol, which was later used in the cell-free system (Fig.8.7) to assess effect of SUZ12 removal on DNA replication. Using protein G magnetic beads and anti-SUZ12 antibody, I subjected the human cytosol to two consecutive rounds of immunodepletion. I carried out a mock immunodepletion using the protein G beads only in parallel with the SUZ12 immunodepletion.

I tested the effectiveness of the immunodepletion on a Western blot (Fig.8.7A). For both the SUZ12 and mock immunodepletions, I loaded the input (untreated cytosol), the immunodepleted cytosol and the eluate from the beads of both immunodepletion rounds, onto a 10% SDS gel. This was developed using the anti-SUZ12 antibody, to ascertain the abundance of the SUZ12 protein in all components.

I tested the resultant immunodepleted cytosol on the human cell-free system to assess the effect of SUZ12 removal on human DNA replication (Fig.8.7B). I substituted human cytosol with the SUZ12- and mock-immunodepleted cytosol and performed a standard replication experiment in the human cell-free system.

Finally, I performed a standard Y RNA depletion of human cytosol for comparison with the immunodepletion study (Fig.8.7C). The cytosol used in the cell-free system underwent a mock depletion (using T3 oligonucleotides) and Y RNA depletion (using complementary Y RNA specific oligonucleotides), prior to being added to a standard replication reaction in the human cell-free system.





**Figure 8.7:** (A) Cytosolic extract of HeLa cells were immunodepleted for two consecutive rounds with magnetic Dynabeads and anti-SUZ12 antibodies (SUZ12) or magnetic beads only (Mock). The input, untreated cytosol, immunodepleted (ID) cytosol (supernatant (SN)) and eluate from both rounds of magnetic beads for both the SUZ12 and Mock conditions were loaded onto a 10% SDS polyacrylamide gel, which was subjected to a standard Western blot protocol and developed with the anti-SUZ12 antibody. Bands for SUZ12 protein, IgG and potentially protein G are indicated on the resultant image. (B) Synchronised late G1 template nuclei were incubated with the SUZ12- or Mock-immunodepleted cytosol and a mixture of dNTPs/NTPs (including a dig-dUTP) in physiological buffer. A negative control where the template nuclei were incubated without cytosol was also conducted (“buffer only”). The nuclei were then spun onto a glass coverslip and stained for DNA and the incorporated dig-dUTP. The percentages of replicating nuclei were found through immunofluorescence microscopy and the mean values, with standard deviations are shown here. A 2-tailed student’s T-test (unequal variance) was performed and demonstrated a significant reduction in replication activity of the SUZ12-immunodepleted cytosol when compared to the Mock-immunodepleted cytosol (n= 3). (C) Synchronised late G1 template nuclei were incubated with the Mock (via incubation with a non-human specific T3 oligonucleotide) and Y RNA (via incubation with a mixture of Y RNA complementary oligonucleotides) depleted cytosol and a mixture of dNTPs/NTPs (including a dig-dUTP) in physiological buffer. A negative control where the template nuclei were incubated without cytosol was also conducted (“buffer only”). The nuclei were then spun onto a glass coverslip and stained for DNA and the incorporated dig-dUTP. The percentages of replicating nuclei were found through immunofluorescence microscopy and the mean values, with standard deviations are shown here. A 2-tailed student’s T-test (unequal variance) was performed and demonstrated a significant reduction in replication activity of the Y RNA- depleted cytosol when compared to the mock-depleted cytosol (n= 4). (\* p < 0.05; \*\* p < 0.01)

The Western blot of the SUZ12 immunodepletion (Fig.8.7A) showed a single band of SUZ12 in the untreated input cytosol. The same single SUZ12 band was present in the immunodepleted cytosol, but abundance was much diminished. The beads of both immunodepletion rounds showed three bands: the first at the molecular weight of SUZ12 (more abundant in the beads from round 1 than round 2); the second was a large band at

approximately 50kDa and was the SUZ12 antibody (IgG - equally abundant in both rounds); the final band present was just above 25kDa in molecular weight (equally abundant in both rounds), which is consistent with the approximate molecular weight of protein G.

The Western blot of the mock (beads only) immunodepletion (Fig.8.7A) showed a single band of the same abundance of SUZ12 in the untreated input cytosol and the immunodepleted cytosol. Both the immunodepletion beads (rounds 1 and 2) showed no SUZ12 band but did have a single band present at just above 25kDa in molecular weight consistent with protein G (bound to the magnetic beads), which was likely to have been eluted from the beads.

Taken together, I conclude that there was a partial immunodepletion of SUZ12 in the SUZ12 immunodepletion, and no immunodepletion of SUZ12 in the mock immunodepletion. Therefore, I was able to use the immunodepleted cytosols in the human cell-free system to assess the effect of the removal of SUZ12 (and any SUZ12 bound proteins) on DNA replication.

Figure 8.7B showed that in the absence of any cytosol, ~10% of nuclei were replicating (indicating ~10% contamination by S-phase nuclei), which increased to approximately 36% replicating nuclei upon the use of the mock immunodepleted cytosol (Mock ID). The replication activity reduced to ~27% when the SUZ12 immunodepleted cytosol (SUZ12 ID) was added, which was a statistically significant reduction compared to the mock ID. Therefore, the removal of SUZ12 by partial immunodepletion leads to significant inhibition of DNA replication.

Figure 8.7C showed the effect of Y RNA depletion on the percentage of replicating nuclei; ~2% of nuclei were replicating in the absence of cytosol. Upon the mock depletion of the cytosol with a T3 oligonucleotide, the percentage of replicating nuclei increased to ~43%. Whereas the depletion of hY RNAs from the cytosol, using Y RNA specific oligonucleotides, resulted in a significant reduction in the percentage of replicating nuclei to ~7%. As observed in previous studies, Y RNAs are required for DNA replication.

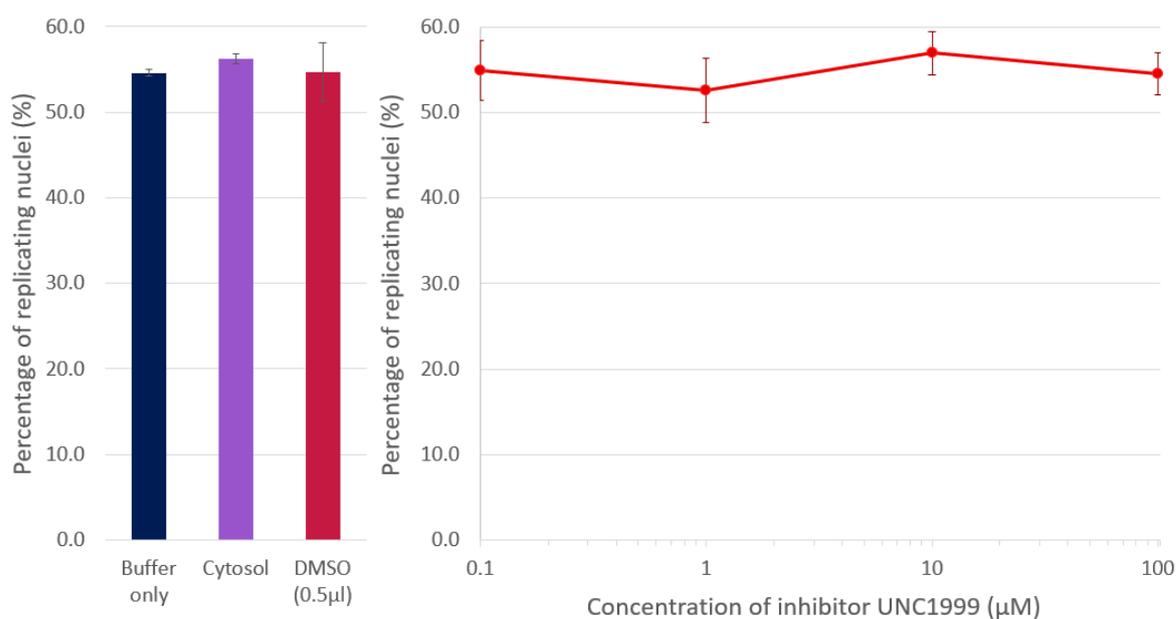
The effect of Y RNA depletion on DNA replication can be compared to that of the immunodepletion and chemical and antibody inhibitions of PRC2 in the human cell-free system. Of the immunodepletion and the chemical inhibition studies, where significant reductions in percentage of replicating nuclei were also observed, the reductions were not as large as that of Y RNA depletion. Where Y RNAs are essential for DNA replication, PRC2 may contribute to DNA replication but these results suggest that it is not the only Y RNA-binding protein complex that is required for DNA replication.

Taken together, these data showed that the removal/inhibition of PRC2 resulted in a partial reduction in DNA replication *in vitro* and I conclude that PRC2 probably plays a stimulatory role in DNA replication but cannot distinguish between DNA replication initiation and elongation, as the cell-free system examines both. I therefore conducted a further study to determine if PRC2 is involved in initiation alone.

#### *8.3.6 Inhibition of PRC2 in S-phase template nuclei*

To assess whether PRC2 plays a role solely in DNA replication initiation (and not elongation), I inhibited the EZH1/2 subunits with the chemical inhibitor UNC1999, in an adaptation of the cell-free system (Fig.8.8). This adaptation, referred to as the elongation cell-free system substitutes late G1-phase nuclei with S-phase synchronised template nuclei. In these nuclei, DNA replication initiation has already taken place and established DNA replication forks are present (111).

Should PRC2 only play a stimulatory role in DNA replication initiation, there would be no effect on replication elongation, and as a result, we would see no effect on the percentage of replicating S-phase nuclei in this elongation cell-free system. If there is a reduction in DNA replication activity in the elongation cell-free system, this would imply a role for PRC2 in replication initiation and/or elongation.



**Figure 8.8:** Synchronised S-phase template nuclei were incubated with untreated, UNC1999-treated (0.1, 1, 10 and 100µM) or DMSO-treated (at the same volume that was used for the chemical inhibitor) cytosolic HeLa extract and a mixture of dNTPs/NTPs (including a dig-dUTP) in physiological buffer. A buffer only control where the template nuclei were incubated without cytosol was also conducted. The nuclei were then spun onto a glass coverslip and stained for DNA and the incorporated dig-dUTP. The percentages of replicating nuclei were found through immunofluorescence microscopy and the mean values, with standard deviations are shown here. A 2-tailed student's T-test (unequal variance) was performed and established no significant difference between all experimental conditions (n= 3).

Figure 8.8 showed that in the absence of cytosol, the percentage of replicating nuclei was ~55%. Upon addition of cytosol, the percentage of replicating nuclei increased to ~56%. Taken together, these controls indicate that the nuclei were successfully synchronised to S-phase.

The addition of DMSO-treated cytosol resulted in no increase in percentage of replicating nuclei compared to the buffer only and cytosol controls, showing that the addition of DMSO to the replication reaction did not impact DNA replication.

Upon the addition of UNC1999-treated cytosol at increasing concentrations of UNC1999 (0.1, 1, 10, 100 µM), the percentage of replicating nuclei for all concentrations were within 52.5% and 56.9% (Fig.8.8). Following a student t-test, the percentage of replicating nuclei of all conditions were not significantly different from the buffer only, cytosol and DMSO controls. These data show that the inhibition of EZH2 and consequently the inhibition of PRC2 had no effect on DNA replication of S-phase nuclei in these replication reactions. It is, therefore, unlikely that PRC2 plays a role in DNA replication elongation.

Taken together, the findings from the inhibition of PRC2 in late G1-phase synchronised nuclei and S-phase synchronised nuclei indicate that the Y RNA-binding chromatin

remodelling complex PRC2 plays a stimulatory role in DNA replication initiation but not in DNA replication elongation.

## **8.4 Discussion**

### *8.4.1 PRC2 is a stimulatory replication initiation factor*

I identified the core subunits and one common cofactor of the PRC2 complex (Table.3.1) from the mass spec. data of Y RNA-binding proteins in a fraction with replication activity (Fig.8.3) (3). My work (Fig.8.4–8.8) showed that PRC2 plays a role in DNA replication initiation but not elongation, which is consistent with Y RNAs' function DNA replication. Unlike Y RNAs, PRC2 appeared to not be essential and only have a stimulatory role in DNA replication initiation.

To establish this stimulatory role for PRC2, I have presented here a series of inhibition studies. With respect to the chemical inhibition study of late G1-phase nuclei (Fig.8.5), EZH1 containing PRC2 complexes have been found to be less affected by SAM inhibitors, including UNC1999. EZH1 containing PRC2 can perform histone methylation, particularly in the absence of EZH2 containing PRC2 (17,19,110). It is possible but unlikely (as UNC1999as concentrations were vastly in excess of the IC<sub>50</sub> concentration (109) for inhibition) that some remaining EZH1 containing PRC2 has escaped inhibition by UNC1999.

The other inhibition studies (antibody and immunodepletion) were of the other minimal core subunits, SUZ12 and EED and showed a similar pattern of partial reduction of replication activity (although not significant in the antibody inhibition) as the chemical inhibition. This further suggests that the partial reduction was unlikely to be due to EZH1 containing PRC2 escaping inhibition.

As stated above, the antibody inhibition of EED and SUZ12 showed a partial but insignificant reduction in DNA replication activity (Fig.8.4). It is possible that the anti-EED and anti-SUZ12 antibodies were not as effective at inhibition of the corresponding proteins whilst they were in their native state; EED and/or SUZ12 may undergo a conformational change, meaning the antibody binding site was unrecognisable, or EED/SUZ12 interacted with additional proteins that obscured the antibody binding site, rendering the antibodies less effective at sequestering their corresponding proteins. The Western blot demonstrated that the antibodies were able to bind the denatured corresponding proteins. Furthermore, the SUZ12 immunodepletion (conducted with the same SUZ12 antibody) required two rounds of immunodepletion to remove SUZ12 sufficiently for testing. There was a small amount of SUZ12 present in the SUZ12-immunodepleted cytosol which may have contributed to the residual replication activity.

Taking these considerations into account, it remains highly likely that PRC2 possesses the previously described stimulatory role in DNA replication initiation.

#### 8.4.2 PRC2 and other chromatin remodellers in DNA replication

As DNA replication is an essential cellular process required for organism viability, it is highly likely that there is redundancy in the process and with Y RNA-binding DNA replication initiation factors (112). We are already aware of 4 functionally redundant Y RNAs that are essential for DNA replication initiation (112). It is probable that, in addition to PRC2, there are further proteins/protein complexes (most likely chromatin remodellers) that both bind to Y RNAs and are also required for DNA replication, potentially explaining why PRC2 has only a stimulatory role.

PRMT1 was identified from the same mass spec. data of Y RNA-binding proteins and is required for DNA replication in the human cell-free system (113). A series of immunodepletion, chemical and antibody inhibition studies showed that PRMT1 was required for DNA replication initiation and elongation. PRMT1 is a methyltransferase that catalyses the transfer of a methyl group from SAM to its target. PRMT1 is the predominant human type I PRMTs that catalyses the asymmetric dimethylation of H4R3 (a marker of transcriptional activity) and is responsible for up to 85% of all arginine methylation. PRMT1 implicated in the DNA damage response, mRNA splicing and epigenetic inheritance (114–119).

At 42kDa, PRMT1 is likely to form part of a complex for it to be consistent with the molecular weight of the active fraction isolated by Kowalski (3). Roberts (113) carried out a PRMT1 pulldown and mass spec. analysis from which I have identified that the only common subunit with PRC2 was RbAp46 (RBBP7).

RbAp46 (RBBP7) and RbAp48 (RBBP4) have been identified in association with several complexes that act in DNA replication. RbAp46/48 is a subunit of the xNuRD complex (a Y RNA-independent DNA replication factor) (47) and the PRC2 complex (4,5). Both RbAp46 and RbAp48 were identified from the mass spec. of Y RNA-binding proteins (3). RbAp46 was also found to interact with PRMT1 (113). To further compound the potential importance of RbAp46/48, SUZ12-Nurf55 (Nurf55 is the *Drosophila* homologue of RbAp46/48) has been established as the minimal component required for nucleosome binding, suggesting that it may contribute to PRC2's high binding affinity for nucleosomes (40). Therefore, RbAp46/48's interactions with Y RNAs may be required for the recruitment of the associated complexes to chromatin at replication origins. However, RbAp46/48 have been found to interact with multiple proteins in other chromatin remodelling complexes (41), where they provide roles in structural support, protein-protein interaction promotion (42–45) and association with nucleosomes (6,40). The recurring appearance of RbAp46/48 may be indicative of an

important role in DNA replication initiation and would warrant further comprehensive and in-depth investigation.

The subunits of the histone chaperone, the FACT complex was also identified from Kowalski's mass spec. data from a Y RNA-binding protein fraction required for DNA replication *in vitro*. Inhibition of the FACT complex also led to a reduction in DNA replication activity in the human cell-free system (T. Krude, Personal Communication).

FACT consists of the two subunits, SUPT16 and SSRP1, and associates with MCM2-7 helicase (120). It is vital for nucleosome reorganisation during transcription, DNA damage signalling and repair, and DNA replication, where SSRP1 depletion results in slowed replication fork progression, but not origin firing (120–122). These suggest a role in elongation, rather than initiation. SUPT16 and SSRP1 were also found in Roberts' PRMT1 pulldown mass spec. data and may also associated with PRMT1 (113).

It is therefore possible that PRC2, PRMT1 and FACT play a combinatorial role in DNA replication. There may also be additional Y RNA-binding proteins that play a role in DNA replication initiation and this requires future investigation.

Additionally, the Y RNA-independent DNA replication factor xNuRD may also have a link with PRC2. NuRD complexes interact with PRC2 and they appear to function antagonistically at some activated genes in embryonic stem cells (123).

#### *8.4.3 PRC2 and RNA binding*

PRC2 is a Y RNA-binding protein (3) but also binds promiscuously with many ncRNAs (96) and I initially considered that there may have been a non-specific interaction with Y RNAs. With the data produced in this work, which demonstrated that the inhibition of PRC2 reduced replication activity, clearly non-specific binding is less likely.

PRC2 binding affinity for RNAs increases as the RNA length increases from 10 to 300 bases. Once the RNA length reaches 300 bases in length, RNA size no longer affects PRC2-RNA binding affinity (96). This finding is highly relevant, as hY RNAs are only 80-120bps in length (112). To assess whether PRC2-Y RNA binding affinity for it is greater than the expected PRC2-RNA binding affinity of the corresponding size, future work could include quantitative electrophoretic mobility shift assay (EMSA) of reconstituted human PRC2, both in the presence and absence of the cofactor AEBP2 (identified in Kowalski's mass spec data) with hY RNAs. This would determine whether PRRC2 has a preferential/specific binding affinity for Y RNAs.

PRC2 have a higher affinity for G-rich regions (97), corresponds with the increased GC content (i.e. GUG-CAG motif and GC-clamps) in the upper stem domain of the hY RNAs (112). PRC2 affinity for RNAs is reduced when RNAs are double stranded (97), suggesting

that the double stranded nature of hY RNAs would lessen their binding affinity. Nevertheless, disruptions to dsRNAs, such as small internal loops and bulges has been found to overcome the reduced binding affinity of dsRNA (97). These features are all present in hY RNAs (112) indicating that their double stranded nature would not affect the binding affinity of PRC2 for hY RNAs. Moreover, SUZ12 has been found to bind to the stem-loop structure of other ncRNAs (37), which implicates it as the potential Y RNA-binding subunit, as Y RNAs possess a stem-loop structure (112).

As yet, no RNA binding site has been identified in the PRC2 complex (89). It may be possible to narrow down which subunit hY RNAs interact with by conducting a binding affinity experiment, whereby each hY RNAs would be added to each individual purified PRC2 subunit (and AEBP2) by EMSAs. However, the individual subunits of PRC2 have altered RNA-binding affinities when they are not in the complex (94,95). There may also be a conformational change in the subunits when they are bound to other subunits, which may increase or decrease their affinity for hY RNAs. These must obviously be taken into consideration when conducting EMSAs.

It may be possible that PRC2 does not interact directly with hY RNAs and an additional 'bridging' protein may be required to facilitate the interaction between hY RNAs and PRC2. This would be similar to the role of RBFox2 that facilitates Rep A binding to PRC2 (100). This should be investigated in future works.

Currently, there are two established models for explaining RNA binding to PRC2: a non-active site inhibitor, and the recruitment of PRC2 to a target site (95,96,124). It is believed that RNA binding inhibits PRC2 from binding to DNA through the induction of a conformational change in PRC2 (96,97).

It is unlikely that Y RNAs inhibit PRC2 as the data presented here showed that inhibition of PRC2 lead to a reduction in DNA replication *in vitro*. If Y RNAs inhibited PRC2, its removal would have increased DNA replication *in vitro*. To confirm that Y RNAs do not inhibit PRC2 activity, future work could include a histone methyltransferase assay of reconstituted human PRC2 in the presence and absence of Y RNAs. If hY RNAs inhibit the function of PRC2, the levels of methylation by PRC2 would be reduced in the presence of hY RNAs when compared to methylation levels in the absence of hY RNAs.

#### 8.4.4 PRC2 and ds-*iniSeq* origins

It is possible that hY RNAs are involved in the recruitment of PRC2 to the genome, potentially to the DNA replication origins. PRC2 predominantly associates at TSS (125,126) and is identified at both transcriptionally active and inactive sites but is not catalytically active where transcription is taking place (94). PRC2 possesses a strong propensity to bind at CpG

islands (101,127,128). TSS and CGIs have been identified as features commonly present at replication origins (129–131). I have confirmed this in my thesis and identified CGIs as the dominant feature correlated with highly efficient DNA replication origin activity (Fig.5.9/12C).

The histone target of PRC2, H3K27, associates with replication origins. H3K27ac associates with early replicating origins/efficient origins (132). H3K27me3 associates with late replicating/less efficient replication origins (133), although H3K27me3 associates with ~40% origins that fire in early–mid S-phase (134). With this literature in mind, PRC2 which trimethylates H3K27 (5) appeared to be a prime candidate for a role in origins specification/activation.

My overlap analysis of ds-*ini*Seq origins with H3K27me3 binding sites (HCT116 ChIP-seq) showed that H3K27me3 sites were almost excluded from ds-*ini*Seq origin sites (Fig.5.13), which implies that PRC2 was not involved in origin specification/activation. It would also suggest that hY RNAs were not required for the recruitment of PRC2 to origins. However, one must always bear in mind that these ChIP-seq data were conducted in a different cell-line and there may be epigenetic differences between HCT116 and EJ30s, resulting in an inaccurate image of the ds-*ini*Seq origins association with H3K27me3. Conversely, I performed the same overlap analysis with 3 other cell-lines (HeLa S3-cervical carcinoma; H1-human embryonic stem cell; PC3-prostate adenocarcinoma), which showed a highly similar pattern (data not shown). This makes the argument of cell-line differences weaker, but ChIP-seq should be conducted in the EJ30 cell-line to negate this argument.

The ds-*ini*Seq origins were found to be present at early replicating sites (Fig.5.6) and H3K27me3 is predominantly associated with origins that fire in late DNA replication (135). It is possible that the lack of association of H3K27me3 with ds-*ini*Seq origins results from replication timing and H3K27me3 is required for the specification and/or activation of the later firing origins that are not identified with ds-*ini*Seq. I was unable to obtain ChIP-seq data of the PRC2 subunits in a sufficiently similar cell-line, which would be ideal data to address the location of PRC2 during DNA replication.

There is interplay between other epigenetic features and PRC2. H2A.Z and H3.3 (associated with early firing origins (136,137)) co-ordinately regulate PRC2-dependent H3K27 trimethylation in mESC (138), and transcriptionally active marks H3K4me3 (also associated with early-replication firing origins) and H3K36me3 strongly inhibit PRC2 (34,86,87). PRC2 may be present at early and late firing origins but remains catalytically inactive at early firing origin sites, as neighbouring histone marks silence PRC2; as seen with transcription (94).

H3K27me3 and H3K4me3 define a poised promoter, which possesses a bivalent state that regulates transcription. The extent of transcription is dependent on the levels of H3K4me3 and H3K27me3, where there is a high occupancy of H3K4me3 at highly active transcription

but there is a lower occupancy of H3K4me3 and high levels of H3K27me3 at less active genes (139,140). This may be applicable to DNA replication.

It is possible that hY RNAs direct PRC2 to DNA replication origin sites, where it does not trimethylate H3K27 at early firing sites which associate with high H3K4me3 levels that inhibit PRC2 but does lay down H3K27me3 at origin sites that fire later in DNA replication. As ds-iniSeq is predominantly comprised of early replication firing origins, this may explain the lack of association between H3K27me3 and ds-iniSeq origins.

#### *8.4.5 Summary*

Overall, I have shown that the Y RNA-binding PRC2 complex plays a stimulatory role in DNA replication initiation but not elongation. The interaction of Y RNAs with PRC2 introduces another chromatin remodeller associated with both Y RNAs and DNA replication and provides a potential mechanism by which Y RNAs fulfil their role in DNA replication initiation.

## **Chapter 9: Conclusion**

In chapter 4, I presented the successful development of a novel NGS method for the identification of human replication origins. I showed that ds-iniSeq is a reliable, reproducible, and viable method, with the ability to determine identified origins' relative replication activities and to subsequently manipulate the *in vitro* replication reaction, upon which this method is based, in order to investigate specific DNA replication factors. The ds-iniSeq origins' replication activities supports the stochastic model of origin firing.

In chapter 5, I performed analysis of the origins identified by the ds-iniSeq method in chapter 4 in greater detail. I found a good concordance with alternative NGS methods used for human origin identification, particularly with origins identified by SNS-seq carried out in the same cell line. I showed that these ds-iniSeq origins were predominantly firing in early replication domains and colocalised with all genomic features associated with human replication origin, with the exception of enhancers. Using the ds-iniSeq origin activities, I was able to highlight CGIs and/or CGI island promoters as dominant features associated with high relative origin activity. Furthermore, I found that the ds-iniSeq origins colocalised strongly with early firing origin-associated histone variant/marks but were almost excluded at late firing origin-associated histone marks. I found that the number of early firing origin-associated histone marks was correlated to ds-iniSeq origin activity; larger numbers of early firing origin associated histone marks colocalised with, and origin was associated with, higher relative origin activity. The ds-iniSeq showed no association with the active transcription mark H3K36me3, thus indicating that there may be a distinction between epigenetic features associated with transcription and DNA replication. Finally, I was able to uncover a correlational link between CGIs/CGI-promoters and early firing origin-associated histone marks, indicating a potential interplay between them.

In chapter 6, I utilised a unique advantage of ds-iniSeq, where I manipulated the *in vitro* replication reaction to investigate the replication initiation factors, Y RNAs and xNuRD. I found that the removal of Y RNAs reduced the number of called origins and their relative origin activities, suggesting a role for Y RNAs in origin activation. Furthermore, I found that the presence of xNuRD did not result in delocalised DNA replication, despite the observation of delocalised DNA replication initiation in pre-MBT *X. laevis* embryos where xNuRD was identified. xNuRD mostly initiated the same ds-iniSeq origins as, and restored relative origin activities to those activated in the presence of Y RNAs, suggesting that Y RNAs and xNuRD predominantly influence the same origin sites and have a potential role on origin activation. The origins influenced by Y RNAs and xNuRD showed similar levels of associations with the genomic and epigenetic features associated with origin firing in early S-phase. I also

identified a small sub-group of new origins that fired in the absence of Y RNAs in later replicating timing windows and possessed unexpectedly high relative origin activities and poor association with genomic and epigenetic features previously colocalised with high origin activity in the presence of Y RNAs and xNuRD. These origins may represent dormant origins that fired as a “failsafe” in the absence of Y RNAs, or origins that normally fire later in replication but activated earlier upon Y RNA removal.

In chapter 7, I introduced the adaptation of the ds-*iniSeq* method in order to investigate replication elongation *in vitro*, which resulted in the development of ds-*eloSeq*. I presented preliminary analysis of the extent of replication elongation away from the origin locations previously identified in ds-*iniSeq*. I found that on average, replication elongation took place in a bidirectional fashion. I showed that the relative origin activities of the ds-*iniSeq* origin locations in the 3-hour ds-*eloSeq* samples were increased in the absence of Y RNAs, but the extent of replication activity away from those origin sites were reduced, when compared to the presence of Y RNAs. These preliminary data require further investigation but may indicate that Y RNAs may not only play a role in origin activation but also in the transition from origin firing to replication elongation/fork progression. I presented the initial development stages of a new method (the Wilkes-Mookerjee (WM) method) for calling discrete origin sites in the 3-hour ds-*eloSeq* data, where it distinguished between the discrete origins and the broader sites of replication elongation away from origins; MACS peak caller was not. The WM method successfully called origin sites in the 3-hour ds-*eloSeq* samples; and a higher proportion of these sites were in mid-late S-phase replication timing windows, when compared to their 15-minute ds-*iniSeq* counterparts. Once the WM method is fully established, it will be an appropriate methodology to assess the number of origins called in other experimental conditions, where Y RNAs were depleted and xNuRD added, and as a basis for identifying DNA replication elongation away from the identified origin sites.

In chapter 8, I investigated the PRC2 complex that was identified from mass spec. data (mass spec. performed by M. Kowalski) of Y RNA-binding proteins. Following a series of inhibition studies using the *in vitro* human cell-free system, I found that PRC2 was a stimulatory factor in DNA replication initiation but not elongation.

Taken together, the work I have presented in this thesis has shown a linked Y RNAs and xNuRD with chromatin remodellers, CGIs/CGI-promoters, epigenetic marks associated with euchromatin, active transcription and origin firing, and higher replication activities identified by ds-*iniSeq*. These data hint at potential mechanisms by which Y RNAs and xNuRD play their roles in DNA replication initiation and origin activation. With these and previous literature in mind, I have hypothesised that Y RNAs interact with chromatin remodellers, including PRC2 and PRMT1, which then potentially targets CGIs/CGI-promoters in order to modify the chromatin environment to establish an environment more conducive with a higher

probability of replication origin firing. Some of my work suggests that xNuRD may fulfil a similar role to Y RNAs. As ever future work must be conducted to establish the validity of my hypothesis.

In summary, the work presented here has contributed to the elucidation of the mechanisms by which Y RNAs and xNuRD fulfil their essential roles in DNA replication initiation, and my subsequent hypothesis has provided a number of avenues for future investigations.

## References

### *Chapter 1: Introduction*

1. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. *Molecular Biology of the Cell* (5th Edition). 2008. Garland Science.
2. Hartwell L, Hood L, Goldberg M, Reynolds A, Silver L. *Genetics: From Genes to Genomes* (4th Edition). 2011. McGraw-Hill.
3. Preston BD, Albertson TM, Herr AJ. DNA Replication Fidelity and Cancer.pdf. *Semin Cancer Biol.* 2011;20(5):281–93.
4. Sancar A, Lindsey-Boltz LA, Ünsal-Kaçmaz K, Linn S. Molecular Mechanisms of Mammalian DNA Repair and the DNA Damage Checkpoints. *Annu Rev Biochem.* 2004;73(1):39–85.
5. Drake JW, Charlesworth B, Charlesworth D, Crow JF. Rates of Spontaneous Mutation *John. Genetics.* 1998;148:1667–86.
6. Fazlieva R, Spittle CS, Morrissey D, Hayashi H, Yan H, Matsumoto Y. Proofreading exonuclease activity of human DNA polymerase  $\delta$  and its effects on lesion-bypass DNA synthesis. *Nucleic Acids Res.* 2009;37(9):2854–66.
7. Mason PA, Cox LS. The role of DNA exonucleases in protecting genome stability and their impact on ageing. *Age (Omaha).* 2012;34(6):1317–40.
8. Bębenek A, Ziuzia-Graczyk I. Fidelity of DNA replication—a matter of proofreading. *Curr Genet [Internet].* 2018;64(5):985–96. Available from: <http://dx.doi.org/10.1007/s00294-018-0820-1>
9. Williams GH, Stoeber K. *The Cell Cycle and Cancer. Med Cell Biol Third Ed.* 2011;(October 2011):273–89.
10. Klug WS, Cummings MR, Spencer CA, Palladino MA. *Concepts of Genetics. Tenth Edit.* Pearson; 2012.
11. Recolin B, van der Laan S, Tsanov N, Maiorano D. Molecular mechanisms of DNA replication checkpoint activation. *Genes (Basel).* 2014;5(1):147–75.
12. Hustedt N, Gasser SM, Shimada K. Replication checkpoint: Tuning and coordination of replication forks in S phase. *Genes (Basel).* 2013;4(3):388–434.

13. Lanz MC, Dibitetto D, Smolka MB. DNA damage kinase signaling: checkpoint and repair at 30 years . *EMBO J.* 2019;38(18).
14. Meselson M, Stahl FW. The replication of DNA in *Escherichia coli*. *Proc Natl Acad Sci U S A.* 1958;44(7):671–82.
15. Bell SP, Dutta A. DNA Replication in Eukaryotic Cells. *Annu Rev Biochem.* 2002;71(1):333–74.
16. Araki H. Elucidating the DDK -dependent step in replication initiation . *EMBO J.* 2016;35(9):907–8.
17. Tanaka S, Tak YS, Araki H. The role of CDK in the initiation step of DNA replication in eukaryotes. *Cell Div.* 2007;2:1–6.
18. O'Donnell M, Langston L, Stillman B. Principles and concepts of DNA replication in bacteria, archaea, and eukarya. *Cold Spring Harb Perspect Biol.* 2013;5(7):1–13.
19. Waga S, Masuda T, Takisawa H, Sugino A. DNA polymerase  $\epsilon$  is required for coordinated and efficient chromosomal DNA replication in *Xenopus* egg extracts. *Proc Natl Acad Sci U S A.* 2001;98(9):4978–83.
20. Fragkos M, Ganier O, Coulombe P, Méchali M. DNA replication origin activation in space and time. *Nat Rev Mol Cell Biol.* 2015;16(6):360–74.
21. Bell SD. Prime-time progress. *Nature.* 2006;439(7076):542–3.
22. Berg J, Tymoczko J, Stryer L. *Biochemistry (5th Edition).* 2002. W H Freeman.
23. Kuhn H, Frank-Kamenetskii MD. Template-independent ligation of single-stranded DNA by T4 DNA ligase. *FEBS J.* 2005;272(23):5991–6000.
24. Craig N, Cohen-Fix O, Green R, Greider C, Storz G, Wolberger C. *Molecular Biology: Principles of Genome Function.* 2010. Oxford University Press.
25. Bailis JM, Forsburg SL. It's all in the timing: linking S phase to chromatin structure and chromosome dynamics. *Cell cycle.* 2003;2(4):302–5.
26. Demeret C, Vassetzky Y, Méchali M. Chromatin remodelling and DNA replication: From nucleosomes to loop domains. *Oncogene.* 2001;20(24 REV. ISS. 3):3086–93.
27. MacAlpine DM, Almouzni G. Chromatin and DNA replication. *Cold Spring Harb Perspect Biol.* 2013;5(8):1–22.
28. Travers A. An engine for nucleosome remodeling. *Cell.* 1999;96(3):311–4.

29. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*. 1997;389(6648):251–60.
30. Budhavarapu VN, Chavez M, Tyler JK. How is epigenetic information maintained through DNA replication? *Epigenetics and Chromatin* [Internet]. 2013;6(1):1. Available from: *Epigenetics & Chromatin*
31. Kouzarides T. Chromatin Modifications and Their Function. *Cell*. 2007;128(4):693–705.
32. Long H, Zhang L, Lv M, Wen Z, Zhang W, Chen X, et al. H2A.Z facilitates licensing and activation of early replication origins. *Nature* [Internet]. 2020;577(7791):576–81. Available from: <http://dx.doi.org/10.1038/s41586-019-1877-9>
33. Goldman MA, Holmquist GP, Gray MC, Caston LA, Nag A. Replication timing of genes and middle repetitive sequences. *Science* (80- ). 1984;224(4650):686–92.
34. Melendy T, Li R. Chromatin remodelling and initiation of DNA replication. *Front Biosci*. 2001;6:1048–53.
35. Bellush JM, Whitehouse I. DNA replication through a chromatin environment. *Philos Trans R Soc B Biol Sci*. 2017;372(1731).
36. Flaus A, Owen-Hughes T. Mechanisms for ATP-dependent chromatin remodelling: The means to the end. *FEBS J*. 2011;278(19):3579–95.
37. Sims JK, Wade PA. Mi-2/NuRD complex function is required for normal S phase progression and assembly of pericentric heterochromatin. *Mol Biol Cell*. 2011;22(17):3094–102.
38. Delamarre A, Barthe A, de la Roche Saint-André C, Luciano P, Forey R, Padioleau I, et al. MRX Increases Chromatin Accessibility at Stalled Replication Forks to Promote Nascent DNA Resection and Cohesin Loading. *Mol Cell*. 2020;77(2):395-410.e3.
39. Bhaskara S. Histone deacetylases 1 and 2 regulate DNA replication and DNA repair: Potential targets for genome stability-mechanism-based therapeutics for a subset of cancers. *Cell Cycle*. 2015;14(12):1779–85.
40. Wilting RH, Yanover E, Heideman MR, Jacobs H, Horner J, Van Der Torre J, et al. Overlapping functions of Hdac1 and Hdac2 in cell cycle regulation and haematopoiesis. *EMBO J* [Internet]. 2010;29(15):2586–97. Available from: <http://dx.doi.org/10.1038/emboj.2010.136>
41. Yamaguchi T, Cubizolles F, Zhang Y, Reichert N, Kohler H, Seiser C, et al. Histone

- deacetylases 1 and 2 act in concert to promote the G1-to-S progression. *Genes Dev* [Internet]. 2009;25:455–69. Available from: [papers3://publication/doi/10.1101/gad.552310](https://pubmed.ncbi.nlm.nih.gov/1911101/)
42. Bhaskara S, Jacques V, Rusche JR, Olson EN, Cairns BR, Chandrasekharan MB. Histone deacetylases 1 and 2 maintain S-phase chromatin and DNA replication fork progression. *Epigenetics and Chromatin*. 2013;6(1).
  43. Clapier CR, Cairns BR. Regulation of ISWI involves inhibitory modules antagonized by nucleosomal epitopes. *Nature*. 2012;492(7428):280–4.
  44. Clapier CR, Cairns BR. The Biology of Chromatin Remodeling Complexes. *Annu Rev Biochem*. 2009;78(1):273–304.
  45. Vincent JA, Kwong TJ, Tsukiyama T. ATP-dependent chromatin remodeling shapes the DNA replication landscape. *Nat Struct Mol Biol*. 2008;15(5):477–84.
  46. Yadav T, Whitehouse I. Replication-coupled nucleosome assembly and positioning by ATP-dependent chromatin remodeling enzymes. *Cell Rep*. 2016;15(4):715–23.
  47. Ehrenhofer-Murray AE. Chromatin dynamics at DNA replication, transcription and repair. *Eur J Biochem*. 2004;271(12):2335–49.
  48. Sporbert A, Gahl A, Ankerhold R, Leonhardt H, Cardoso MC. DNA polymerase clamp shows little turnover at established replication sites but sequential de novo assembly at adjacent origin clusters. *Mol Cell*. 2002;10(6):1355–65.
  49. Shibahara KI, Stillman B. Replication-dependent marking of DNA by PCNA facilitates CAF-1-coupled inheritance of chromatin. *Cell*. 1999;96(4):575–85.
  50. Stewart-Morgan KR, Petryk N, Groth A. Chromatin replication and epigenetic cell memory. *Nat Cell Biol* [Internet]. 2020;22(4):361–71. Available from: <http://dx.doi.org/10.1038/s41556-020-0487-y>
  51. Li Z, Hua X, Serra-Cardona A, Xu X, Gan S, Zhou H, et al. DNA polymerase  $\alpha$  interacts with H3-H4 and facilitates the transfer of parental histones to lagging strands. *Sci Adv*. 2020;6(35).
  52. Symeonidou IE, Taraviras S, Lygerou Z. Control over DNA replication in time and space. *FEBS Lett* [Internet]. 2012;586(18):2803–12. Available from: <http://dx.doi.org/10.1016/j.febslet.2012.07.042>
  53. Tsakraklides V, Bell SP. Dynamics of pre-replicative complex assembly. *J Biol Chem*. 2010;285(13):9437–43.

54. Klemm RD, Austin RJ, Bell SP. Coordinate binding of ATP and origin DNA regulates the ATPase activity of the origin recognition complex. *Cell*. 1997;88(4):493–502.
55. Chesnokov IN. Multiple Functions of the Origin Recognition Complex. *Int Rev Cytol*. 2007;256(March):69–109.
56. Takara TJ, Bell SP. Multiple Cdt1 molecules act at each origin to load replication-competent Mcm2-7 helicases. *EMBO J* [Internet]. 2011;30(24):4885–96. Available from: <http://dx.doi.org/10.1038/emboj.2011.394>
57. Kang S, Kang MS, Ryu E, Myung K. Eukaryotic DNA replication: Orchestrated action of multi-subunit protein complexes. *Mutat Res - Fundam Mol Mech Mutagen* [Internet]. 2018;809(May 2017):58–69. Available from: <https://doi.org/10.1016/j.mrfmmm.2017.04.002>
58. Bell SP, Labib K. Chromosome duplication in *Saccharomyces cerevisiae*. *Genetics*. 2016;203(3):1027–67.
59. Blow JJ. Control of chromosomal DNA replication in the early *Xenopus* embryo. *EMBO J*. 2001;20(13):3293–7.
60. Thome KC, Dhar SK, Quintana DG, Delmolino L, Shahsafaei A, Dutta A. Subsets of human origin recognition complex (ORC) subunits are expressed in non-proliferating cells and associate with non-ORC proteins. *J Biol Chem*. 2000;275(45):35233–41.
61. Coster G, Frigola J, Beuron F, Morris E, Diffley J. Origin Licensing Requires ATP Binding and Hydrolysis by the MCM Replicative Helicase. *Mol Cell* [Internet]. 2014;55(5):667–77. Available from: <http://dx.doi.org/10.1016/j.biochi.2015.03.025><http://dx.doi.org/10.1038/nature10402><http://dx.doi.org/10.1038/nature21059><http://journal.stainkudus.ac.id/index.php/equilibrium/article/view/1268/1127><http://dx.doi.org/10.1038/nrmicro2577><http://>
62. Sun J, Evrin C, Samel SA, Fernández-Cid A, Riera A, Kawakami H, et al. Cryo-EM structure of a helicase loading intermediate containing ORC-Cdc6-Cdt1-MCM2-7 bound to DNA. *Nat Struct Mol Biol*. 2013;20(8):944–51.
63. Popova V V., Brechalov A V., Georgieva SG, Kopytova D V. Nonreplicative functions of the origin recognition complex. *Nucleus* [Internet]. 2018;9(1):460–73. Available from: <https://doi.org/10.1080/19491034.2018.1516484>
64. Vashee S, Cvetic C, Lu W, Simancek P, Kelly TJ, Walter JC. Sequence-independent DNA binding and replication initiation by the human origin recognition complex. *Genes*

- Dev. 2003;17(15):1894–908.
65. Okayama H. Cdc6: A trifunctional AAA+ ATPase that plays a central role in controlling the G1-S transition and cell survival. *J Biochem.* 2012;152(4):297–303.
  66. Borlado LR, Méndez J. CDC6: From DNA replication to cell cycle checkpoints and oncogenesis. *Carcinogenesis.* 2008;29(2):237–43.
  67. Piatti S, Böhm T, Cocker JH, Diffley JFX, Nasmyth K. Activation of S-phase-promoting CDKs in late G1 defines a “point of no return” after which Cdc6 synthesis cannot promote DNA replication in yeast. *Genes Dev.* 1996;10(12):1516–31.
  68. Drury LS, Perkins G, Diffley JFX. The Cdc4/34/53 pathway targets Cdc6p for proteolysis in budding yeast. *EMBO J.* 1997;16(19):5966–76.
  69. Sánchez M, Calzada A, Bueno A. The Cdc6 protein is ubiquitinated in vivo for proteolysis in *Saccharomyces cerevisiae*. *J Biol Chem.* 1999;274(13):9092–7.
  70. Speck C, Chen Z, Li H, Stillman B. ATPase-dependent, cooperative binding of ORC and Cdc6p to origin DNA. *Nat Struct Mol Biol.* 2005;12(11):965–71.
  71. Weinreich M, Liang C, Stillman B. The Cdc6p nucleotide-binding motif is required for loading Mcm proteins onto chromatin. *Proc Natl Acad Sci U S A.* 1999;96(2):441–6.
  72. Coleman TR, Carpenter PB, Dunphy WG. The *Xenopus* Cdc6 protein is essential for the initiation of a single round of DNA replication in cell-free extracts. *Cell.* 1996;87(1):53–63.
  73. Mahadevappa R, Neves H, Yuen SM, Bai Y, McCrudden CM, Yuen HF, et al. The prognostic significance of Cdc6 and Cdt1 in breast cancer. *Sci Rep.* 2017;7(985):1–11.
  74. Maiorano D, Moreau J, Méchali M. XCDT1 is required for the assembly of pre-replicative complexes in *Xenopus laevis*. *Nature.* 2000;404(6778):622–5.
  75. Sugimoto N, Tatsumi Y, Tsurumi T, Matsukage A, Kiyono T, Nishitani H, et al. Cdt1 Phosphorylation by Cyclin A-dependent Kinases Negatively Regulates Its Function without Affecting Geminin Binding. *J Biol Chem.* 2004;279(19):19691–7.
  76. You Z, Masai H. Cdt1 forms a complex with the minichromosome maintenance protein (MCM) and activates its helicase activity. *J Biol Chem.* 2008;283(36):24469–77.
  77. Pozo P, Cook J. Regulation and Function of Cdt1; A Key Factor in Cell Proliferation and Genome Stability. *Genes (Basel).* 2016;8(1):2.

78. Wohlschlegel JA, Dwyer BT, Dhar SK, Cvetic C, Walter JC, Dutta A. Inhibition of eukaryotic DNA replication by geminin binding to Cdt1. *Science* (80- ). 2000;290(5500):2309–12.
79. Vijayraghavan S, Schwacha A. The Eukaryotic Mcm2-7 Replicative Helicase. *Subcell Biochem.* 2012;62:113–34.
80. Sato M, Gotow T, You Z, Komamura-Kohno Y, Uchiyama Y, Yabuta N, et al. Electron microscopic observation and single-stranded DNA binding activity of the Mcm4,6,7 complex. *J Mol Biol.* 2000;300(3):421–31.
81. Ishimi Y. A DNA helicase activity is associated with an MCM4, -6, and -7 protein complex. *J Biol Chem.* 1997;272(39):24508–13.
82. Bochman ML, Bell SP, Schwacha A. Subunit Organization of Mcm2-7 and the Unequal Role of Active Sites in ATP Hydrolysis and Viability. *Mol Cell Biol.* 2008;28(19):5865–73.
83. You Z, Komamura Y, Ishimi Y. Biochemical Analysis of the Intrinsic Mcm4-Mcm6-Mcm7 DNA Helicase Activity. *Mol Cell Biol.* 1999;19(12):8003–15.
84. Labib K, Gambus A. A key role for the GINS complex at DNA replication forks. *Trends Cell Biol.* 2007;17(6):271–8.
85. Shen Z, Prasanth SG. Emerging players in the initiation of eukaryotic DNA replication. *Cell Div.* 2012;7(1):22.
86. Ilves I, Petojevic T, Pesavento JJ, Botchan MR, Biology C. Activation of the MCM2-7 Helicase by Association with Cdc45 and GINS proteins. *Mol Cell.* 2010;37(2):247–58.
87. Gómez-Escoda B, Wu P-YJ. Roles of CDK and DDK in genome duplication and maintenance: meiotic singularities. *Genes (Basel).* 2017;8(3):8–12.
88. Marks AB, Fu H, Aladjem MI. Regulation of replication origins. *Adv Exp Med Biol.* 2017;1042:43–59.
89. Takeda DY, Dutta A. DNA replication and progression through S phase. *Oncogene.* 2005;24(17):2827–43.
90. Masai H, Taniyama C, Ogino K, Matsui E, Kakusho N, Matsumoto S, et al. Phosphorylation of MCM4 by Cdc7 kinase facilitates its interaction with Cdc45 on the chromatin. *J Biol Chem [Internet].* 2006;281(51):39249–61. Available from: <http://dx.doi.org/10.1074/jbc.M608935200>

91. Furstenthal L, Kaiser BK, Swanson C, Jackson PK. Cyclin E uses Cdc6 as a chromatin-associated receptor required for DNA replication. *J Cell Biol.* 2001;152(6):1267–78.
92. Boos D, Sanchez-Pulido L, Rappas M, Pearl LH, Oliver AW, Ponting CP, et al. Regulation of DNA replication through Sld3-Dpb11 interaction is conserved from yeast to humans. *Curr Biol [Internet].* 2011;21(13):1152–7. Available from: <http://dx.doi.org/10.1016/j.cub.2011.05.057>
93. Kang YH, Galal WC, Farina A, Tappin I, Hurwitz J. Properties of the human Cdc45/Mcm2-7/GINS helicase complex and its action with DNA polymerase  $\epsilon$  in rolling circle DNA synthesis. *Proc Natl Acad Sci U S A.* 2012;109(16):6042–7.
94. Mueller AC, Keaton MA, Dutta A. DNA replication: Mammalian treslin-topBP1 interaction mirrors yeast Sld3-Dpb11. *Curr Biol [Internet].* 2011;21(16):R638–40. Available from: <http://dx.doi.org/10.1016/j.cub.2011.07.004>
95. Kumagai A, Shevchenko A, Shevchenko A, Dunphy WG. Direct regulation of Treslin by cyclin-dependent kinase is essential for the onset of DNA replication. *J Cell Biol.* 2011;193(6):995–1007.
96. Heller RC, Kang S, Lam WM, Chen S, Chan CS, Bell SP. Eukaryotic Origin-Dependent DNA Replication in vitro Reveals Sequential Action of DDK and S-CDK Kinases. *Cell.* 2011;146(1):80–91.
97. Reuswig KU, Pfander B. Control of eukaryotic DNA replication initiation—Mechanisms to ensure smooth transitions. *Genes (Basel).* 2019;10(2).
98. Sanuki Y, Kubota Y, Kanemaki MT, Takahashi TS, Mimura S, Takisawa H. RecQ4 promotes the conversion of the pre-initiation complex at a site-specific origin for DNA unwinding in xenopus egg extracts. *Cell Cycle.* 2015;14(7):1010–23.
99. Pospiech H, Grosse F, Pisani F. The Initiation Step of Eukaryotic DNA Replication. *Subcell Biochem.* 2010;50:79–104.
100. Lee KY, Bang SW, Yoon SW, Lee S-H, Yoon J-B, Hwang DS. Phosphorylation of ORC2 Protein Dissociates Origin Recognition Complex from Chromatin and Replication origins. *J Biol Chem.* 2012;287(15):11891–8.
101. Burgers PMJ, Kunkel TA. Eukaryotic DNA Replication Fork. *Annu Rev Biochem.* 2017 Jun 20;86(1):417–38.
102. Li H, O'Donnell M. The eukaryotic CMG helicase at the replication fork: Emerging

- architecture reveals an unexpected mechanism. *Bioessays*. 2018;40(3):1–19.
103. Szambowska A, Tessmer I, Prus P, Schlott B, Pospiech H, Grosse F. Cdc45-induced loading of human RPA onto single-stranded DNA. *Nucleic Acids Res*. 2017;45(6):3217–30.
  104. Mimura S, Takisawa H. *Xenopus* Cdc45-dependent loading of DNA polymerase  $\alpha$  onto chromatin under the control of S-phase cdk. *EMBO J*. 1998;17(19):5699–707.
  105. Takaya J, Kusunoki S, Ishimi Y. Protein interaction and cellular localization of human CDC45. *J Biochem*. 2013;153(4):381–8.
  106. Köhler C, Koalick D, Fabricius A, Parplys AC, Borgmann K, Pospiech H, et al. Cdc45 is limiting for replication initiation in humans. *Cell Cycle*. 2016;15(7):974–85.
  107. Carroni M, De March M, Medagli B, Krastanova I, Taylor IA, Amenitsch H, et al. New insights into the GINS complex explain the controversy between existing structural models. *Sci Rep* [Internet]. 2017;7(August 2016):1–7. Available from: <http://dx.doi.org/10.1038/srep40188>
  108. Aparicio T, Ibarra A, Méndez J. Cdc45-MCM-GINS, a new power player for DNA replication. *Cell Div*. 2006;1(Mcm):4–6.
  109. Choi JM, Lim HS, Kim JJ, Song O-K, Cho Y. Crystal structure of the human GINS complex. *Genes Dev*. 2007;21(11):1316–21.
  110. Lööke M, Maloney MF, Bell SP. Mcm10 regulates DNA replication elongation by stimulating the CMG replicative helicase. *Genes Dev*. 2017;31(3):291–305.
  111. Moreno SP, Gambus A. Mechanisms of eukaryotic replisome disassembly. *Biochem Soc Trans*. 2020;48(3):823–36.
  112. Fien K, Hurwitz J. Fission yeast Mcm10p contains primase activity. *J Biol Chem*. 2006;281(31):22248–60.
  113. Quan Y, Xia Y, Liu L, Cui J, Li Z, Cao Q, et al. Cell-Cycle-Regulated Interaction between Mcm10 and Double Hexameric Mcm2-7 Is Required for Helicase Splitting and Activation during S Phase. *Cell Rep* [Internet]. 2015;13(11):2576–86. Available from: <http://dx.doi.org/10.1016/j.celrep.2015.11.018>
  114. Kawasaki Y, Hiraga SI, Sugino A. Interactions between Mcm10p and other replication factors are required for proper initiation and elongation of chromosomal DNA replication in *Saccharomyces cerevisiae*. *Genes to Cells*. 2000;5(12):975–89.

115. Wohlschlegel JA, Dhar SK, Prokhorova TA, Dutta A, Walter JC. Xenopus Mcm10 binds to origins of DNA replication after Mcm2-7 and stimulates origin binding of Cdc45. *Mol Cell*. 2002;9(2):233–40.
116. Simon AC, Zhou JC, Perera RL, Van Deursen F, Evrin C, Ivanova ME, et al. A Ctf4 trimer couples the CMG helicase to DNA polymerase  $\delta$  in the eukaryotic replisome. *Nature*. 2014;510(7504):293–7.
117. Ricke RM, Bielinsky AK. Mcm10 regulates the stability and chromatin association of DNA polymerase- $\alpha$ . *Mol Cell*. 2004;16(2):173–85.
118. O'Donnell M, Li H. The Eukaryotic Replisome Goes Under the Microscope. *Curr Biol*. 2016;26(6):R247–56.
119. Cotterill S, Kearsey S. Eukaryotic DNA polymerases. *Encycl Life Sci John Wiley Sons*. 2009;77–87.
120. Schneider A, Smith RWP, Kautz AR, Weissbart K, Grosse F, Nasheuer HP. Primase activity of human DNA polymerase  $\alpha$ -primase. *J Biol Chem*. 1998;273(34):21608–15.
121. Frick DN, Richardson CC. DNA Primases. *Annu Rev Biochem*. 2001;70:39–80.
122. Liu T, Huang J. Replication protein A and more: Single-stranded DNA-binding proteins in eukaryotic cells. *Acta Biochim Biophys Sin (Shanghai)*. 2016;48(7):665–70.
123. Zou Y, Liu Y, Wu X, Shell SM. Replication to DNA Damage and Stress Responses. *J Cell Physiol*. 2006;208(2):267–73.
124. Iftode C, Daniely Y, Borowiec JA. Replication protein A (RPA): The eukaryotic SSB. *Crit Rev Biochem Mol Biol*. 1999;34(3):141–80.
125. Deng SK, Chen H, Symington LS. Replication Protein A prevents promiscuous annealing between short sequence homologies: Implications for genome integrity. *Bioessays*. 2015;37(3):305–13.
126. Kowalski MP, Krude T. Functional roles of non-coding Y RNAs. *Int J Biochem Cell Biol* [Internet]. 2015;66:20–9. Available from: <http://dx.doi.org/10.1016/j.biocel.2015.07.003>
127. Christov CP, Gardiner TJ, Szuts D, Krude T. Functional Requirement of Noncoding Y RNAs for Human Chromosomal DNA Replication. *Mol Cell Biol*. 2006;26(18):6993–7004.
128. Jacob F, Brenner S, Cuzin F. On the regulation of the initiation of DNA replication in bacteria. *Cold Spring Harb Symp Quant Biol*. 1968;28:329–48.

129. Prioleau MN, MacAlpine DM. DNA replication origins—Where do we begin? *Genes Dev.* 2016;30(15):1683–97.
130. Rowley A, Cocker JH, Harwood J, Diffley JF. Initiation complex assembly at budding yeast replication origins begins with the recognition of a bipartite sequence by limiting amounts of the initiator, ORC. *EMBO J.* 1995;14(11):2631–41.
131. Besnard E, Babled A, Lapasset L, Milhavet O, Parrinello H, Dantec C, et al. Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat Struct Mol Biol.* 2012;19(8):837–44.
132. Cayrou C, Ballester B, Peiffer I, Fenouil R, Coulombe P, Andrau JC, et al. The chromatin environment shapes DNA replication origin organization and defines origin classes. *Genome Res.* 2015;25(12):1873–85.
133. Martin MM, Ryan M, Kim RG, Zakas AL, Fu H, Lin CM, et al. Genome-wide depletion of replication initiation events in highly transcribed regions. *Genome Res.* 2011;21(11):1822–32.
134. Smith OK, Kim R, Fu H, Martin MM, Lin CM, Utani K, et al. Distinct epigenetic features of differentiation-regulated replication origins. *Epigenetics and Chromatin.* 2016;9(1):1–17.
135. Kitsberg D, Selig S, Keshet I, Cedar H. Replication structure of the human p-globin gene domain D. *Lett to Nat.* 1993;366:588–90.
136. Shima N, Pederson KD. Dormant origins as a built-in safeguard in eukaryotic DNA replication against genome instability and disease development Naoko. *DNA Repair (Amst).* 2017;56:166–73.
137. Sugimoto N, Maehara K, Yoshida K, Ohkawa Y, Fujita M. Genome-wide analysis of the spatiotemporal regulation of firing and dormant replication origins in human cells. *Nucleic Acids Res.* 2018;46(13):6683–96.
138. Picard F, Cadoret JC, Audit B, Arneodo A, Alberti A, Battail C, et al. The Spatiotemporal Program of DNA Replication Is Associated with Specific Combinations of Chromatin Marks in Human Cells. *PLoS Genet.* 2014;10(5).
139. Foulk MS, Urban JM, Casella C, Gerbi SA. Characterizing and controlling intrinsic biases of lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-quadruplex motifs around a subset of human replication origins. *Genome Res.* 2015;25(5):725–35.

140. Fu H, Besnard E, Desprat R, Ryan M, Kahli M, Lemaitre J-M, et al. Mapping Replication Origin Sequences in Eukaryotic Chromosomes. *Curr Protoc Cell Biol.* 2001;65(1):1–7.
141. Sequeira-Mendes J, Díaz-Uriarte R, Apedaile A, Huntley D, Brockdorff N, Gómez M. Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genet.* 2009;5(4).
142. Cadoret JC, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, Duret L, et al. Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc Natl Acad Sci U S A.* 2008;105(41):15837–42.
143. Mesner LD, Valsakumar V, Cieslik M, Pickin M, Hamlin JL, Bekiranov S. Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early- and late-firing origins. *Genome Res.* 2013;23(11):1774–88.
144. Mesner LD, Crawford EL, Hamlin JL. Isolating apparently pure libraries of replication origins from complex genomes. *Mol Cell.* 2006;21(5):719–26.
145. Mesner LD, Valsakumar V, Karnani N, Dutta A, Hamlin JL, Bekiranov S. Bubble-chip analysis of human origin distributions demonstrates on a genomic scale significant clustering into zones and significant association with transcription (*Genome Research* (2011) 21, (377-389)). *Genome Res.* 2011;21(9):1561.
146. Petryk N, Kahli M, D'Aubenton-Carafa Y, Jaszczyszyn Y, Shen Y, Silvain M, et al. Replication landscape of the human genome. *Nat Commun.* 2016;7:1–13.
147. Smith DJ, Whitehouse I. Intrinsic coupling of lagging-strand synthesis to chromatin assembly. *Nature* [Internet]. 2006;483(7390):434–8. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf>
148. Langley AR, Gräf S, Smith JC, Krude T. Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). *Nucleic Acids Res.* 2016;44(21):10230–47.
149. Marchal C, Sasaki T, Vera D, Wilson K, Sima J, Rivera-Mulia JC, et al. Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq. *Nat Protoc* [Internet]. 2018;13(5):819–39. Available from: <http://dx.doi.org/10.1038/nprot.2017.148>
150. Zhao PA, Sasaki T, Gilbert DM. High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome Biol.* 2020;21(1):76.

151. Meldi L, Brickner JH. Compartmentalization of the nucleus. *Trends Cell Biol.* 2011;21(12):701–8.
152. Duan Z, Blau CA. The genome in space and time: Does form always follow function? *BioEssays.* 2012;34(9):800–10.
153. Rhind N, Gilbert DM. DNA replication timing. *Cold Spring Harb Perspect Biol.* 2013;5(8):1–26.
154. Rivera-Mulia JC, Gilbert DM. Replication timing and transcriptional control: beyond cause and effect—part III. *Curr Opin Cell Biol.* 2016;40:168–78.
155. Di Tomaso MV, Liddle P, Lafon-Huges L, Reyes-Abalos AL, Folle G. Chromatin Damage Patterns Shift According to Eu/ Heterochromatin Replication. In: *The Mechanisms of DNA Replication* [Internet]. 2013. p. DOI: 10.5772/51847. Available from: <http://dx.doi.org/10.1016/j.tws.2012.02.007>
156. Jost KL, Bertulat B, Cardoso MC. Heterochromatin and gene positioning: Inside, outside, any side? *Chromosoma.* 2012;121(6):555–63.
157. Tamaru H. Confining euchromatin/heterochromatin territory: Jumonji crosses the line. *Genes Dev.* 2010;24(14):1465–78.
158. Imai R, Nozaki T, Tani T, Kaizu K, Hibino K, Ide S, et al. Density imaging of heterochromatin in live cells using orientation-independent-DIC microscopy. *Mol Biol Cell.* 2017;28(23):3349–59.
159. Cremer T, Cremer C. CHROMOSOME TERRITORIES, NUCLEAR ARCHITECTURE AND GENE REGULATION IN MAMMALIAN CELLS. *Nat Rev Genet.* 2001;2:292–301.
160. Hiratani I, Takahashi S. DNA replication timing enters the single-cell era. *Genes (Basel).* 2019;10(3).
161. Gindin Y, Valenzuela MS, Aladjem MI, Meltzer PS, Bilke S. A chromatin structure-based model accurately predicts DNA replication timing in human cells. *Mol Syst Biol.* 2014;10(3).
162. Madrigal P, Krajewski P. Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. *Front Genet.* 2012;3(OCT):1–3.
163. Bass HW, Hoffman GG, Lee TJ, Wear EE, Joseph SR, Allen GC, et al. Defining multiple, distinct, and shared spatiotemporal patterns of DNA replication and endoreduplication from 3D image analysis of developing maize (*Zea mays* L.) root tip

- nuclei. *Plant Mol Biol*. 2015;89(4–5):339–51.
164. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. *Nature*. 2017;485(7398):376–80.
  165. van Steensel B, Belmont AS. Lamina-associated domains: links with chromosome architecture, heterochromatin and gene repression. *Cell*. 2017;169(5):780–91.
  166. Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, et al. Topologically-associating domains are stable units of replication-timing regulation. *Nature*. 2014;515(7527):402–5.
  167. Navendra V, Rocha P, An D, Raviram R, Skok JA, Mazzoni EO, et al. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* (80- ). 2015;347(6225):1017–21.
  168. Cayrou C, Coulombe P, Puy A, Rialle S, Kaplan N, Segal E, et al. New insights into replication origin characteristics in metazoans. *Cell Cycle*. 2012;11(4):658–67.
  169. Pherson M, Misulovin Z, Gause M, Dorsett D. Cohesin occupancy and composition at enhancers and promoters are linked to DNA replication origin proximity in *Drosophila*. *Genome Res*. 2019;29(4):602–12.
  170. Rhodes D, Lipps HJ. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res*. 2015;43(18):8627–37.
  171. Gardiner-Garden M, Frommer M. CpG Islands in Vertebrate Genomes. *J Mol Biol*. 1987;196:261–82.
  172. Delgado S, Gómez M, Bird A, Antequera F. Initiation of DNA replication at CpG islands in mammalian chromosomes. *EMBO J*. 1998;17(8):2426–35.
  173. Prioleau MN. CpG Islands: Starting blocks for replication and transcription. *PLoS Genet*. 2009;5(4):4–5.
  174. Long HK, King HW, Patient RK, Odom DT, Klose RJ. Protection of CpG islands from DNA methylation is DNA-encoded and evolutionarily conserved. *Nucleic Acids Res*. 2016;44(14):6693–706.
  175. Lombrana R, Almeida R, Álvarez A, Gómez M. R-loops and initiation of DNA replication in human cells: A missing link? *Front Genet*. 2015;6(APR):1–7.
  176. Valton AL, Hassan-Zadeh V, Lema I, Boggetto N, Alberti P, Saintomé C, et al. G4

- motifs affect origin positioning and efficiency in two vertebrate replicators. *EMBO J.* 2014;33(7):732–46.
177. Prorok P, Artufel M, Aze A, Coulombe P, Peiffer I, Lacroix L, et al. Involvement of G-quadruplex regions in mammalian replication origin activity. *Nat Commun* [Internet]. 2019;10(1):1–16. Available from: <http://dx.doi.org/10.1038/s41467-019-11104-0>
178. Valton AL, Prioleau MN. G-Quadruplexes in DNA Replication: A Problem or a Necessity? *Trends Genet* [Internet]. 2016;32(11):697–706. Available from: <http://dx.doi.org/10.1016/j.tig.2016.09.004>
179. Hassan-Zadeh V, Chilaka S, Cadoret JC, Ma MKW, Boggetto N, West AG, et al. USF binding sequences from the HS4 insulator element impose early replication timing on a vertebrate replicator. *PLoS Biol.* 2012;10(3).
180. Keller H, Kiosze K, Sachsenweger J, Haumann S, Ohlenschläger O, Nuutinen T, et al. The intrinsically disordered amino-terminal region of human RecQL4: multiple DNA-binding domains confer annealing, strand exchange and G4 DNA binding. *Nucleic Acids Res.* 2014;42(20):12614–27.
181. Hoshina S, Yura K, Teranishi H, Kiyasu N, Tominaga A, Kadoma H, et al. Human origin recognition complex binds preferentially to G-quadruplex-preferable RNA and single-stranded DNA. *J Biol Chem.* 2013;288(42):30161–71.
182. De Magis A, Manzo SG, Russo M, Marinello J, Morigi R, Sordet O, et al. DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells. *Proc Natl Acad Sci U S A.* 2019;116(3):816–25.
183. Skourti-Stathaki K, Proudfoot NJ. A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. *Genes Dev* [Internet]. 1984;28:1384–96. Available from: <http://www.genesdev.org/cgi/doi/10.1101/gad.242990.114>.
184. Chen Y-H, Keegan S, Kahli M, Tonzi P, Fenyo D, Huang TT, et al. Transcription shapes DNA replication initiation and termination in human cells. *Nat Struct Mol Biol.* 2019;26(1):67–77.
185. Fu H, Aladjem MI. Replication timing and nuclear structure. *Curr Opin Cell Biol* [Internet]. 2018;52:43–50. Available from: <file:///C:/Users/Carla Carolina/Desktop/Artigos para acrescentar na qualificação/The impact of birth weight on cardiovascular disease risk in the.pdf>
186. Aladjem MI. Replication in context: Dynamic regulation of DNA replication patterns in

- metazoans. *Nat Rev Genet.* 2007;8(8):588–600.
187. Dailey L. High throughput technologies for the functional discovery of mammalian enhancers: New approaches for understanding transcriptional regulatory network dynamics. *Genomics* [Internet]. 2015;106(3):151–8. Available from: <http://dx.doi.org/10.1016/j.ygeno.2015.06.004>
188. Andersson R, Sandelin A. Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet* [Internet]. 2020;21(2):71–87. Available from: <http://dx.doi.org/10.1038/s41576-019-0173-8>
189. Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr ARW, James KD, Turner DJ, et al. Orphan CpG Islands Identify numerous conserved promoters in the mammalian genome. *PLoS Genet.* 2010;6(9).
190. Hamperl S, Cimprich KA. Conflict Resolution in the Genome: How Transcription and Replication Make It Work. *Cell.* 2016;167(6):1455–67.
191. Helmrich A, Ballarino M, Tora L. Collisions between Replication and Transcription Complexes Cause Common Fragile Site Instability at the Longest Human Genes. *Mol Cell* [Internet]. 2011;44(6):966–77. Available from: <http://dx.doi.org/10.1016/j.molcel.2011.10.013>
192. Gates LA, Foulds CE, O'Malley BW. Histone marks in the 'drivers seat': functional roles in steering the transcription cycle. *Trends Biochem Sci.* 2017;42(12):977–89.
193. Hiratani I, Takebayashi S ichiro, Lu J, Gilbert DM. Replication timing and transcriptional control: beyond cause and effect-part II. *Curr Opin Genet Dev.* 2009;19(2):142–9.
194. Casas-Delucchi CS, Van Bommel JG, Haase S, Herce HD, Nowak D, Meilinger D, et al. Histone hypoacetylation is required to maintain late replication timing of constitutive heterochromatin. *Nucleic Acids Res.* 2012;40(1):159–69.
195. Unnikrishnan A, Gafken PR, Tsurimoto T. Dynamic changes in histone acetylation regulate origins of DNA replication. *Nat Struct Mol Biol.* 2010;17(4):430–7.
196. Vogelauer M, Rubbi L, Lucas I, Brewer BJ, Grunstein M. Histone Acetylation Regulates the Time of Replication Origin Firing. *Mol Cell* [Internet]. 2002;10(5):1223–33. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12453428>
197. Goren A, Tabib A, Hecht M, Cedar H. DNA replication timing of the human  $\beta$ -globin domain is controlled by histone modification at the origin. *Genes Dev.*

- 2008;22(10):1319–24.
198. Rhind N. DNA replication timing: Random thoughts about origin firing. *Nat Cell Biol.* 2006;8(12):1313–6.
  199. Rhind N, Yang SCH, Bechhoefer J. Reconciling stochastic origin firing with defined replication timing. *Chromosom Res.* 2010;18(1):35–43.
  200. Local A, Huang H, Albuquerque CP, Singh N, Lee AY, Wang W, et al. Identification of H3K4me1-associated proteins at mammalian enhancers. *Nat Genet* [Internet]. 2018;50(1):73–82. Available from: <http://dx.doi.org/10.1038/s41588-017-0015-6>
  201. Fu H, Maunakea AK, Martin MM, Huang L, Zhang Y, Ryan M, et al. Methylation of Histone H3 on Lysine 79 Associates with a Group of Replication Origins and Helps Limit DNA Replication Once per Cell Cycle. *PLoS Genet.* 2013;9(6):1–14.
  202. Iizuka M, Matsui T, Takisawa H, Smith MM. Regulation of Replication Licensing by Acetyltransferase Hbo1. *Mol Cell Biol.* 2006;26(3):1098–108.
  203. Struhl K, Miotto B. HBO1 histone acetylase activity is essential for DNA replication licensing and inhibited by geminin. *Mol Cell.* 2010;37(January 2010):211–20.
  204. Kuo AJ, Song J, Cheung P, Ishibe-Murakami S, Yamazoe S, Chen JK, et al. ORC1 BAH domain links H4K20 to DNA replication licensing and Meier-Gorlin syndrome. *Nature* [Internet]. 2012;484(7392):115–9. Available from: <http://dx.doi.org/10.1038/nature10956>
  205. Brustel J, Kirstein N, Izard F, Grimaud C, Prorok P, Cayrou C, et al. Histone H4K20 trimethylation at late-firing origins ensures timely heterochromatin replication. *EMBO J.* 2017;36(18):2726–41.
  206. Lubelsky Y, Prinz JA, DeNapoli L, Li Y, Belsky JA, MacAlpine DM. DNA replication and transcription programs respond to the same chromatin cues. *Genome Res.* 2014;24(7):1102–14.
  207. Doyon Y, Cayrou C, Ullah M, Landry AJ, Côté V, Selleck W, et al. ING tumor suppressor proteins are critical regulators of chromatin acetylation required for genome expression and perpetuation. *Mol Cell.* 2006;21(1):51–64.
  208. Costas C, De M, Sanchez P, Stroud H, Yu Y, Oliveros C, et al. Genome-wide mapping of Arabidopsis origins of DNA replication and their associated epigenetic marks. *Nat Struct Mol Biol.* 2011;18(3):395–400.

209. Ruan K, Yamamoto TG, Asakawa H, Chikashige Y, Kimura H, Masukata H, et al. Histone H4 acetylation required for chromatin decompaction during DNA replication. *Sci Rep.* 2015;5:1–10.
210. Iizuka M, Stillman B. Histone acetyltransferase HBO1 interacts with the ORC1 subunit of the human initiator protein. *J Biol Chem.* 1999;274(33):23027–34.
211. Burke TW, Cook JG, Asano M, Nevins JR. Replication Factors MCM2 and ORC1 Interact with the Histone Acetyltransferase HBO1. *J Biol Chem.* 2001;276(18):15397–408.
212. Feng Y, Vlassis A, Roques C, Lalonde M, González-Aguilera C, Lambert J, et al. BRPF 3- HBO 1 regulates replication origin activation and histone H3K14 acetylation. *EMBO J.* 2016;35(2):176–92.
213. Yan K, You L, Degerny C, Ghorbani M, Liu X, Chen L, et al. The Chromatin Regulator BRPF3 Preferentially Activates the HBO1 Acetyltransferase but Is Dispensable for Mouse Development and Survival. *J Biol Chem.* 2016;291(6):2647–63.
214. Dykhuizen EC, Hargreaves DC, Miller E, Cui K, Korshunov A, Kool M, et al. mSWI/SNF (BAF) Complexes Facilitate Decatenation of DNA by Topoisomerase II $\alpha$ . *Nature.* 2013;497(7451):624–7.
215. Chammas P, Mocavini I, Di Croce L. Engaging chromatin: PRC2 structure meets function. *Br J Cancer [Internet].* 2019;2(September). Available from: <http://dx.doi.org/10.1038/s41416-019-0615-2>
216. Hałasa M, Wawruszak A, Przybyszewska A, Jaruga A, Guz M, Kałafut J, et al. H3K18Ac as a Marker of Cancer Progression and Potential Target of Anti-Cancer Therapy. *Cells.* 2019;8(5):485.
217. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet.* 2008;40(7):897–903.
218. Wood K, Tellier M, Murphy S. DOT1L and H3K79 methylation in transcription and genomic stability. *Biomolecules.* 2018;8(1):1–16.
219. Tardat M, Brustel J, Kirsh O, Lefevbre C, Callanan M, Sardet C, et al. The histone H4 Lys 20 methyltransferase PR-Set7 regulates replication origins in mammalian cells. *Nat Cell Biol.* 2010;12(11):1086–93.
220. Beck DB, Burton A, Oda H, Ziegler-Birling C, Torres-Padilla ME, Reinberg D. The role

- of PR-Set7 in replication licensing depends on Suv4-20h. *Genes Dev.* 2012;26(23):2580–9.
221. Abbas T, Shibata E, Park J, Jha S, Karnani N, Dutta A. CRL4Cdt2 Regulates Cell Proliferation and Histone Gene Expression by Targeting PR-Set7/Set8 for Degradation Tarek. *Mol Cell* [Internet]. 2011;40(1):9–21. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf>
222. Yang H, Pesavento JJ, Starnes TW, Cryderman DE, Wallrath LL, Kelleher NL, et al. Preferential dimethylation of histone H4 lysine 20 by Suv4-20. *J Biol Chem.* 2008;283(18):12085–92.
223. Pesavento JJ, Yang H, Kelleher NL, Mizzen CA. Certain and Progressive Methylation of Histone H4 at Lysine 20 during the Cell Cycle. *Mol Cell Biol.* 2008;28(1):468–86.
224. Luense LJ, Wang X, Schon SB, Weller AH, Lin Shiao E, Bryant JM, et al. Comprehensive analysis of histone post-translational modifications in mouse and human male germ cells. *Epigenetics and Chromatin.* 2016;9(1):1–15.
225. Kylie K, Romero J, Lindamulage IKS, Knockleby J, Lee H. Dynamic regulation of histone H3K9 is linked to the switch between replication and transcription at the Dbf4 origin-promoter locus. *Cell Cycle* [Internet]. 2016;15(17):2321–35. Available from: <http://dx.doi.org/10.1080/15384101.2016.1201254>
226. Wu R, Wang Z, Zhang H, Gan H, Zhang Z. H3K9me3 demethylase Kdm4d facilitates the formation of pre-initiative complex and regulates DNA replication. *Nucleic Acids Res.* 2017;45(1):169–80.
227. Siefert J, Georgescu C, Wren JD, Koren A, Sansam CL. DNA Replication Timing During Development Anticipates Transcriptional Programs and Parallels Enhancer Activation. *Genome Res.* 2017;27(8):1406–16.
228. Creighton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A.* 2010;107(50):21931–6.
229. Sen S, Block KF, Pasini A, Baylin SB, Easwaran H. Genome-wide positioning of bivalent mononucleosomes. *BMC Med Genomics* [Internet]. 2016;9(1):1–14. Available from: <http://dx.doi.org/10.1186/s12920-016-0221-6>
230. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, et al. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell.* 2006;125(2):315–26.

231. Clément C, Orsi GA, Gatto A, Boyarchuk E, Forest A, Hajj B, et al. High-resolution visualization of H3 variants during replication reveals their controlled recycling. *Nat Commun* [Internet]. 2018;9(1). Available from: <http://dx.doi.org/10.1038/s41467-018-05697-1>
232. Hatch CL, Bonner WM. Sequence of CDNAs for mammalian H2A.Z, an evolutionarily diverged but highly conserved basal histone H2A isoprotein species. *Nucleic Acids Res.* 1988;16(3):1113–24.
233. Lin CJ, Conti M, Ramalho-Santos M. Histone variant H3.3 maintains a decondensed chromatin state essential for mouse preimplantation development. *Dev.* 2013;140(17):3624–34.
234. Chen P, Zhao J, Wang Y, Wang M, Long H, Liang D, et al. H3.3 actively marks enhancers and primes gene transcription via opening higher-ordered chromatin. *Genes Dev.* 2013;27(19):2109–24.
235. Krude T. Initiation of chromosomal DNA replication in mammalian cell-free systems. *Cell Cycle.* 2006;5(18):2115–22.
236. Krude T. Initiation of human DNA replication in vitro using nuclei from cells arrested at an initiation-competent state. *J Biol Chem.* 2000;275(18):13699–707.
237. Marheineke K, Hyrien O, Krude T. Visualization of bidirectional initiation of chromosomal DNA replication in a human cell free system. *Nucleic Acids Res.* 2005;33(21):6931–41.
238. Krude T, Jackman M, Pines J, Laskey RA. Cyclin/Cdk-dependent initiation of DNA replication in a human cell-free system. *Cell* [Internet]. 1997;88(1):109–19. Available from: [http://dx.doi.org/10.1016/S0092-8674\(00\)81863-2](http://dx.doi.org/10.1016/S0092-8674(00)81863-2)
239. Christov CP, Dingwell KS, Skehel M, Wilkes HS, Sale JE, Smith JC, et al. A NuRD Complex from *Xenopus laevis* Eggs Is Essential for DNA Replication during Early Embryogenesis. *Cell Rep* [Internet]. 2018;22(9):2265–78. Available from: <https://doi.org/10.1016/j.celrep.2018.02.015>
240. Mattick JS, Makunin I V. Non-coding RNA. *Hum Mol Genet.* 2006;15 Spec No(1):17–29.
241. Cobb M. Who discovered messenger RNA? *Curr Biol* [Internet]. 2015;25(13):R526–32. Available from: <http://dx.doi.org/10.1016/j.cub.2015.05.032>
242. Cech TR, Steitz JA. The noncoding RNA revolution - Trashing old rules to forge new

- ones. *Cell* [Internet]. 2014;157(1):77–94. Available from:  
<http://dx.doi.org/10.1016/j.cell.2014.03.008>
243. Giral H, Landmesser U, Kratzer A. Into the Wild: GWAS Exploration of Non-coding RNAs. *Front Cardiovasc Med*. 2018;5(December).
244. Wang J, Samuels DC, Zhao S, Xiang Y, Zhao YY, Guo Y. Current research on non-coding ribonucleic acid (RNA). *Genes (Basel)*. 2017;8(12).
245. Pertea M. The human transcriptome: An unfinished story. *Genes (Basel)*. 2012;3(3):344–60.
246. Qu H, Fang X. A Brief Review on the Human Encyclopedia of DNA Elements (ENCODE) Project. *Genomics, Proteomics Bioinforma* [Internet]. 2013;11(3):135–41. Available from: <http://dx.doi.org/10.1016/j.gpb.2013.05.001>
247. Kung JTY, Colognori D, Lee JT. Long noncoding RNAs: Past, present, and future. *Genetics*. 2013;193(3):651–69.
248. Dieci G, Preti M, Montanini B. Eukaryotic snoRNAs: A paradigm for gene expression flexibility. *Genomics* [Internet]. 2009;94(2):83–8. Available from:  
<http://dx.doi.org/10.1016/j.ygeno.2009.05.002>
249. Hendrick J, Wolin S, Rinke J, Lerner M, Steitz J. Ro Small Cytoplasmic Ribonucleoproteins are a subclass of La Ribonucleoproteins. *Mol Cell Biol*. 1981;1(12):1138–49.
250. Lerner MR, Boyle JA, Hardin JA, Streitz JA. Two novel classes of small ribonucleoproteins detected by antibodies associated with Lupus Erythematosus. *Science (80- )*. 1981;211(January):400–2.
251. van Venrooij WJ, Slobbe RL, Pruijn GJM. Structure and function of La and Ro RNPs. *Mol Biol Rep*. 1993;18(2):113–9.
252. Dhahbi JM, Spindler SR, Atamna H, Boffelli D, Martin DIK. Deep Sequencing of Serum Small RNAs Identifies Patterns of 5' tRNA Half and YRNA Fragment Expression Associated with Breast Cancer. *Biomark Cancer*. 2014;6:BIC.S20764.
253. Hall AE, Dalmy T. Discovery of novel small RNAs in the quest to unravel genome complexity. *Biochem Soc Trans*. 2013;41(4):866–70.
254. Köhn M, Pazaitis N, Hüttelmaier S. Why Y RNAs? about versatile RNAs and their functions. *Biomolecules*. 2013;3(1):143–56.

255. Nicolas FE, Hall AE, Csorba T, Turnbull C, Dalmay T. Biogenesis of y RNA-derived small RNAs is independent of the microRNA pathway. *FEBS Lett* [Internet]. 2012;586(8):1226–30. Available from: <http://dx.doi.org/10.1016/j.febslet.2012.03.026>
256. Godoy PM, Bhakta NR, Barczak AJ, Cakmak H, Fisher S, Mackenzie TC, et al. Large Differences in Small RNA Composition Between Human Biofluids. *Cell Rep*. 2018;25(5):1346–58.
257. Driedonks TAP, Nolte-T'Hoën ENM. Circulating Y-RNAs in extracellular vesicles and ribonucleoprotein complexes; Implications for the immune system. *Front Immunol*. 2019;10(JAN):1–15.
258. Kabeerdoss J, Sandhya P, Danda D. Y RNA derived small RNAs in Sjögren's syndrome: Candidate biomarkers? *Int J Rheum Dis*. 2017;20(11):1763–6.
259. Liu YM, Tseng CH, Chen YC, Yu WY, Ho MY, Ho CY, et al. Exosome-delivered and y RNA-derived small RNA suppresses influenza virus replication. *J Biomed Sci*. 2019;26(1):1–14.
260. Cambier L, Couto G, Ibrahim A, Echavez AK, Valle J, Liu W, et al. Y RNA fragment in extracellular vesicles confers cardioprotection via modulation of IL -10 expression and secretion . *EMBO Mol Med*. 2017;9(3):337–52.
261. Farris AD, Gross JK, Hanas JS, Harley JB. Genes for murine Y1 and Y3 Ro RNAs have class 3 RNA polymerase III promoter structures and are unlinked on mouse chromosome 6. *Gene*. 1996;174(1):35–42.
262. Mosig A, Guofeng M, Stadler BMR, Stadler PF. Evolution of the vertebrate Y RNA cluster. *Theory Biosci*. 2007;126(1):9–14.
263. O'Brien CA, Margelot K, Wolin SL. Xenopus Ro ribonucleoproteins: Members of an evolutionarily conserved class of cytoplasmic ribonucleoproteins. *Proc Natl Acad Sci U S A*. 1993;90(15):7250–4.
264. Perreault J, Perreault JP, Boire G. Ro-associated Y RNAs in metazoans: Evolution and diversification. *Mol Biol Evol*. 2007;24(8):1678–89.
265. Maraia R, Sakulich AL, Brinkmann E, Green ED. Gene encoding human Ro-associated autoantigen Y5 RNA. *Nucleic Acids Res*. 1996;24(18):3552–9.
266. Maraia RJ, Sasaki-tozawa N, Driscoll CT, Green ED, Darlington GJ. The human Y4 small cytoplasmic RNA gene is controlled by upstream elements and resides on chromosome 7 with all other hY scRNA genes. *Nucleic Acids Res*. 1994;22(15):3045–

52.

267. Teunissen SWM. Conserved features of Y RNAs: a comparison of experimentally derived secondary structures. *Nucleic Acids Res.* 2000;28(2):610–9.
268. van Gelder CW g., Thijssen JP h. m., Klaassen ECJ, Sturchler C, Krol A, Van Walther J, et al. Common structural features of the Ro RNP associated hY1 and hY5 RNAs. *Nucleic Acids Res.* 1994;22(13):2498–506.
269. Wolin SL, Steitz JA. The Ro small cytoplasmic ribonucleoproteins: Identification of the antigenic protein and its binding site on the Ro RNAs. *Proc Natl Acad Sci U S A.* 1984;81(7 I):1996–2000.
270. Pruijn GJM, Siobbe RL, Venrooij WJ va. Analysis of protein-RNA interactions within Ro ribonucleoprotein complexes. *Nucleic Acids Res.* 1991;19(19):5173–80.
271. Green CD, Long KS, Shi H, Wolin SL. Binding of the 60-kDa Ro autoantigen to Y RNAs: Evidence for recognition in the major groove of a conserved helix. *Rna.* 1998;4(7):750–65.
272. Boccitto M, Wolin SL. Ro60 and Y RNAs: structure, functions, and roles in autoimmunity. *Crit Rev Biochem Mol Biol [Internet].* 2019;54(2):133–52. Available from: <https://doi.org/10.1080/10409238.2019.1608902>
273. Sim S, Wolin SL. Emerging roles for the Ro 60-kDa autoantigen in noncoding RNA metabolism. *Wiley Interdiscip Rev RNA.* 2011;2(5):686–99.
274. Stein AJ, Fuchs G, Fu C, Wolin SL, Reinisch KM. Structural insights into RNA quality control: The Ro autoantigen binds misfolded RNAs via its central cavity. *Cell.* 2005;121(4):529–39.
275. Chen X, Smith JD, Shi H, Yang DD, Flavell RA, Wolin SL. The Ro Autoantigen Binds Misfolded U2 Small Nuclear RNAs and Assists Mammalian Cell Survival after UV Irradiation. *Curr Biol.* 2003;13(24):2206–11.
276. Labbe J-C, Hekimi S, Rokeach L. The levels of the RoRNP-associated Y RNA are dependent upon the presence of ROP-1, the *Caenorhabditis elegans* Ro60 protein. *Genetics [Internet].* 1999;151:143–50. Available from: <http://www.wormbase.org/db/misc/paper?name=WBPaper00003364>
277. O'Brien CA, Wolin SL. A possible role for the 60-kD Ro autoantigen in a discard pathway for defective 5S rRNA precursors. *Genes Dev.* 1994;8(23):2891–903.
278. Fuchs G, Stein AJ, Fu C, Reinisch KM, Wolin SL. Structural and biochemical basis for

- misfolded RNA recognition by the Ro autoantigen. *Nat Struct Mol Biol.* 2006;13(11):1002–9.
279. MacRae IJ, Doudna JA. Ro's role in RNA reconnaissance. *Cell* [Internet]. 2005;121(4):495–6. Available from: <http://dx.doi.org/10.1016/j.cell.2005.05.004>
280. Sim S, Weinberg DE, Fuchs G, Choi K, Chung J, Wolin SL. The Subcellular Distribution of an RNA Quality Control Protein, the Ro Autoantigen, Is Regulated by Noncoding Y RNA Binding. *Mol Biol Cell.* 2009;20:2673–83.
281. Wolin SL, Cedervall T. The La protein. *Annu Rev Biochem.* 2002;71:375–403.
282. Stefano JE. Purified lupus antigen Ia recognizes an oligouridylate stretch common to the 3' termini of RNA polymerase III transcripts. *Cell.* 1984;36(1):145–54.
283. Belisova A, Semrad K, Mayer O, Kocian G, Waigmann E, Schroeder R, et al. RNA chaperone activity of protein components of human Ro RNPs. *Rna.* 2005;11(7):1084–94.
284. Langley AR, Chambers H, Christov CP, Krude T. Ribonucleoprotein particles containing non-coding Y RNAs, Ro60, Ia and nucleolin are not required for Y RNA function in DNA replication. *PLoS One.* 2010;5(10).
285. Farris AD, Koelsch G, Pruijn GJM, Van Venrooij WJ, Harley JB. Conserved features of Y RNAs revealed by automated phylogenetic secondary structure analysis. *Nucleic Acids Res.* 1999;27(4):1070–8.
286. Zhang AT, Langley AR, Christov CP, Kheir E, Shafee T, Gardiner TJ, et al. Dynamic interaction of Y RNAs with chromatin and initiation proteins during human DNA replication. *J Cell Sci.* 2011;124(12):2058–69.
287. Fabini G, Raijmakers R, Hayer S, Fouraux MA, Pruijn GJM, Steiner G. The Heterogeneous Nuclear Ribonucleoproteins I and K Interact with a Subset of the Ro Ribonucleoprotein-associated Y RNAs in Vitro and in Vivo. *J Biol Chem.* 2001;276(23):20711–8.
288. Fouraux MA, Bouvet P, Verkaart S, Van Venrooij WJ, Pruijn GJM. Nucleolin associates with a subset of the human Ro ribonucleoprotein complexes. *J Mol Biol.* 2002;320(3):475–88.
289. Bouffard P, Barbar E, Brière F, Boire G. Interaction cloning and characterization of RoBPI, a novel protein binding to human Ro ribonucleoproteins. *Rna.* 2000;6(1):66–78.

290. Gardiner TJ, Christov CP, Langley AR, Krude T. A conserved motif of vertebrate Y RNAs essential for chromosomal DNA replication. *Rna*. 2009;15(7):1375–85.
291. Wang I, Kowalski MP, Langley AR, Rodriguez R, Balasubramanian S, Hsu STD, et al. Nucleotide Contributions to the Structural Integrity and DNA Replication Initiation Activity of Noncoding y RNA. *Biochemistry*. 2014;53(37):5848–63.
292. Krude T, Christov CP, Hyrien O, Marheineke K. Y RNA functions at the initiation step of mammalian chromosomal DNA replication. *J Cell Sci*. 2009;122(16):2836–45.
293. Christov CP, Trivier E, Krude T. Noncoding human Y RNAs are overexpressed in tumours and required for cell proliferation. *Br J Cancer*. 2008;98(5):981–8.
294. Collart C, Christov CP, Smith JC, Krude T. The Midblastula Transition Defines the Onset of Y RNA-Dependent DNA Replication in *Xenopus laevis*. *Mol Cell Biol*. 2011;31(18):3857–70.
295. Farris AD, Puvion-Dutilleul F, Puvion E, Harley JB, Lee LA. The ultrastructural localization of 60-kDa Ro protein and human cytoplasmic RNAs: Association with novel electron-dense bodies. *Proc Natl Acad Sci U S A*. 1997;94(7):3040–5.
296. Gendron M, Roberge D, Boire G. Heterogeneity of human Ro ribonucleoproteins (RNPS): Nuclear retention of Ro RNPS containing the human hY5 RNA in human and mouse cells. *Clin Exp Immunol*. 2001;125(1):162–8.
297. Hall AE, Turnbull C, Dalmay T. Y RNAs: Recent developments. *Biomol Concepts*. 2013;4(2):103–10.
298. Kheir E, Krude T. Non-coding Y RNAs associate with early replicating euchromatin in concordance with the origin recognition complex. *J Cell Sci*. 2017;130(7):1239–50.
299. Hogg JR, Collins K. Human Y5 RNA specializes a Ro ribonucleoprotein for 5S ribosomal RNA quality control. *Genes Dev*. 2007;21(23):3067–72.
300. Tatsumi Y, Ohta S, Kimura H, Tsurimoto T, Obuse C. The ORC1 cycle in human cells: I. Cell cycle-regulated oscillation of human ORC1. *J Biol Chem*. 2003;278(42):41528–34.
301. Chen X, Quinn AM, Wolin SL. Ro ribonucleoproteins contribute to the resistance of *Deinococcus radiodurans* to ultraviolet irradiation. *Genes Dev*. 2000;14(7):777–82.
302. Chen X, Sim S, Wurtmann EJ, Feke A, Wolin SL. Bacterial noncoding Y RNAs are widespread and mimic tRNAs. *Rna*. 2014;20(11):1715–24.

303. Chen X, Taylor DW, Fowler CC, Galan JE, Wang H-W, Wolin SL. An RNA Degradation Machine Sculpted by Ro Autoantigen and Noncoding RNA. *Cell*. 2013;158(1):166–77.
304. Lima Neto QA de, Junior FFD, Bueno PSA, Seixas FAV, Kowalski MP, Kheir E, et al. Structural and functional analysis of four non-coding Y RNAs from Chinese hamster cells: Identification, molecular dynamics simulations and DNA replication initiation assays. *BMC Mol Biol*. 2016;17(1):1–9.
305. Kowalski MP, Baylis HA, Krude T. Non-coding stem-bulge RNAs are required for cell proliferation and embryonic development in *C. elegans*. *J Cell Sci*. 2015;128(11):2118–29.
306. Duarte Junior FF, Lima Neto QA De, Rando FDS, Freitas DVB De, Pattaro Júnior JR, Polizelli LG, et al. Identification and molecular structure analysis of a new noncoding RNA, a sbRNA homolog, in the silkworm *Bombyx mori* genome. *Mol Biosyst*. 2015;11(3):801–8.
307. Duarte Junior FF, Bueno PSA, Pedersen SL, Rando F dos S, Pattaro Júnior JR, Caligari D, et al. Identification and characterization of stem-bulge RNAs in *Drosophila melanogaster*. *RNA Biol* [Internet]. 2019;16(3):330–9. Available from: <https://doi.org/10.1080/15476286.2019.1572439>
308. Perreault J, Perreault J, Boire G. Ro-Associated Y RNAs in Metazoans : Evolution and Diversification. *Mol Biol Evol*. 2007;24(8):1678–89.
309. Sim S, Wolin SL. Bacterial Y RNAs: Gates, Tethers and tRNA Mimics. *Microbiol Spectr*. 2018;6(4):1–21.
310. Newport J, Kirschner M. A major developmental transition in early xenopus embryos: I. characterization and timing of cellular changes at the midblastula stage. *Cell*. 1982;30(3):675–86.
311. Lemaitre J-M, Geraud G, Mechali M. Dynamics of the Genome during Early *Xenopus laevis* Development: Karyomeres As Independent Units of Replication. *J Cell Biol*. 1998;142(5):1159–66.
312. Mahbubani HM, Paull T, Eider JK, Blow JJ. DNA replication initiates at multiple sites on plasmid DNA in xenopus egg extracts. *Nucleic Acids Res*. 1992;20(7):1457–62.
313. Blow JJ, Gillespie PJ, Francis D, Jackson DA. Replication origins in *Xenopus* egg extract are 5-15 kilobases apart and are activated in clusters that fire at different times. *J Cell Biol*. 2001;152(1):15–25.

314. Wade PA, Jones PL, Vermaak D, Wolffe AP. A multiple subunit Mi-2 histone deacetylase from *Xenopus laevis* cofractionates with an associated Snf2 superfamily ATPase. *Curr Biol.* 1998;8(14):843–8.
315. Denslow SA, Wade PA. The human Mi-2/NuRD complex and gene regulation. *Oncogene.* 2007;26(37):5433–8.
316. Pfefferli C, Müller F, Jazwińska A, Wicky C. Specific NuRD components are required for fin regeneration in zebrafish. *BMC Biol.* 2014;12:1–17.
317. Passannante M, Marti CO, Pfefferli C, Moroni PS, Kaeser-Pebernard S, Puoti A, et al. Different Mi-2 complexes for various developmental functions in *Caenorhabditis elegans*. *PLoS One.* 2010;5(10).
318. Lai AY, Wade PA. NuRD: A multi-faceted chromatin remodeling complex in regulating cancer biology. *Nat Rev Cancer.* 2011;11(8):588–96.
319. Millard CJ, Varma N, Saleh A, Morris K, Watson PJ, Bottrill AR, et al. The structure of the core NuRD repression complex provides insights into its interaction with chromatin. *Elife.* 2016;5(APRIL2016):1–21.
320. Hoffmeister H, Fuchs A, Erdel F, Pinz S, Gröbner-Ferreira R, Bruckmann A, et al. CHD3 and CHD4 form distinct NuRD complexes with different yet overlapping functionality. *Nucleic Acids Res.* 2017;45(18):10534–54.
321. Allen HF, Wade PA, Kutateladze TG. The NuRD architecture. *Cell Mol Life Sci.* 2013;70(19):3513–24.
322. Lodomery M, Lyons S, Sommerville J. *Xenopus* HDm, a maternally expressed histone deacetylase, belongs to an ancient family of acetyl-metabolizing enzymes. *Gene.* 1997;198(2 JAN):275–80.
323. Ryan J, Llinas AJ, White DA, Turner BM, Sommerville J. Maternal histone deacetylase is accumulated in the nuclei of *Xenopus* oocytes as protein complexes with potential enzyme activity. *J Cell Sci.* 1999;112(14):2441–52.
324. Langley AR, Smith JC, Stemple DL, Harvey SA. New insights into the maternal to zygotic transition. *Dev.* 2014;141(20):3834–41.

### *Chapter 3: Materials and Methods*

1. Krude T. Initiation of human DNA replication in vitro using nuclei from cells arrested at an initiation-competent state. *J Biol Chem.* 2000;275(18):13699–707.
2. Krude T, Jackman M, Pines J, Laskey RA. Cyclin/Cdk-dependent initiation of DNA replication in a human cell-free system. *Cell [Internet].* 1997;88(1):109–19. Available from: [http://dx.doi.org/10.1016/S0092-8674\(00\)81863-2](http://dx.doi.org/10.1016/S0092-8674(00)81863-2)
3. Christov CP, Gardiner TJ, Szuts D, Krude T. Functional Requirement of Noncoding Y RNAs for Human Chromosomal DNA Replication. *Mol Cell Biol.* 2006;26(18):6993–7004.
4. Christov CP, Dingwell KS, Skehel M, Wilkes HS, Sale JE, Smith JC, et al. A NuRD Complex from *Xenopus laevis* Eggs Is Essential for DNA Replication during Early Embryogenesis. *Cell Rep [Internet].* 2018;22(9):2265–78. Available from: <https://doi.org/10.1016/j.celrep.2018.02.015>
5. Langley AR, Gräf S, Smith JC, Krude T. Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). *Nucleic Acids Res.* 2016;44(21):10230–47.
6. Petryk N, Kahli M, D'Aubenton-Carafa Y, Jaszczyszyn Y, Shen Y, Silvain M, et al. Replication landscape of the human genome. *Nat Commun.* 2016;7:1–13.
7. Mesner LD, Valsakumar V, Cieslik M, Pickin M, Hamlin JL, Bekiranov S. Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early- and late-firing origins. *Genome Res.* 2013;23(11):1774–88.
8. Dellino GI, Cittaro D, Piccioni R, Luzi L, Banfi S, Segalla S, et al. Genome-wide mapping of human DNA-replication origins: Levels of transcription at ORC1 sites regulate origin selection and replication timing. *Genome Res.* 2013;23(1):1–11.

### *Chapter 4: The development of density-substitution initiation-site sequencing (ds-iniSeq)*

1. Langley AR, Gräf S, Smith JC, Krude T. Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). *Nucleic Acids Res.* 2016;44(21):10230–47.

2. Meselson M, Stahl FW. The replication of DNA in Escherichia coli. Proc Natl Acad Sci U S A. 1958;44(7):671–82.
3. Szybalski W. [124] Use of cesium sulfate for equilibrium density gradient centrifugation. In: Methods in Enzymology [Internet]. 1968. p. 330–60. Available from: [https://doi.org/10.1016/0076-6879\(67\)12149-6](https://doi.org/10.1016/0076-6879(67)12149-6)
4. SeqMonk. Babraham Bioinformatics Seqmonk Project [Internet]. Vol. Version 1. [cited 2020 Sep 20]. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/>
5. Romiguier J, Ranwez V, Douzery EJP, Galtier N. Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. Genome Res. 2010;20(8):1001–9.
6. Panijpan B. The buoyant density of DNA and the G + C content. J Chem Educ. 1977;54(3):172–3.
7. Schildkraut CL, Marmur J, Doty P. Determination of the base composition of deoxyribonucleic acid from its buoyant density in CsCl. J Mol Biol [Internet]. 1962;4(6):430–43. Available from: [http://dx.doi.org/10.1016/S0022-2836\(62\)80100-4](http://dx.doi.org/10.1016/S0022-2836(62)80100-4)
8. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):1–9.
9. Zhao PA, Sasaki T, Gilbert DM. High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. Genome Biol. 2020;21(1):76.
10. Vassilev L, Johnson EM. An initiation zone of chromosomal DNA replication located upstream of the c-myc gene in proliferating HeLa cells. Mol Cell Biol. 1990;10(9):4899–904.

*Chapter 5: The analysis of the standard density-substitution initiation-sites sequencing (ds-*iniSeq*) experiment*

1. Langley AR, Gräf S, Smith JC, Krude T. Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (*ini-seq*). Nucleic Acids Res. 2016;44(21):10230–47.
2. Besnard E, Babled A, Lapasset L, Milhavet O, Parrinello H, Dantec C, et al. Unraveling cell type-specific and reprogrammable human replication origin signatures

- associated with G-quadruplex consensus motifs. *Nat Struct Mol Biol.* 2012;19(8):837–44.
3. Popova V V., Brechalov A V., Georgieva SG, Kopytova D V. Nonreplicative functions of the origin recognition complex. *Nucleus* [Internet]. 2018;9(1):460–73. Available from: <https://doi.org/10.1080/19491034.2018.1516484>
  4. Miotto B, Ji Z, Struhl K. Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers. *Proc Natl Acad Sci U S A.* 2016;113(33):E4810–9.
  5. Picard F, Cadoret JC, Audit B, Arneodo A, Alberti A, Battail C, et al. The Spatiotemporal Program of DNA Replication Is Associated with Specific Combinations of Chromatin Marks in Human Cells. *PLoS Genet.* 2014;10(5).
  6. Cayrou C, Ballester B, Peiffer I, Fenouil R, Coulombe P, Andrau JC, et al. The chromatin environment shapes DNA replication origin organization and defines origin classes. *Genome Res.* 2015;25(12):1873–85.
  7. Petryk N, Kahli M, D'Aubenton-Carafa Y, Jaszczyszyn Y, Shen Y, Silvain M, et al. Replication landscape of the human genome. *Nat Commun.* 2016;7:1–13.
  8. Hyrien O. Peaks cloaked in the mist: The landscape of mammalian replication origins. *J Cell Biol.* 2015;208(2):147–60.
  9. Mesner LD, Valsakumar V, Cieslik M, Pickin M, Hamlin JL, Bekiranov S. Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early- and late-firing origins. *Genome Res.* 2013;23(11):1774–88.
  10. O'Donnell M, Langston L, Stillman B. Principles and concepts of DNA replication in bacteria, archaea, and eukarya. *Cold Spring Harb Perspect Biol.* 2013;5(7):1–13.
  11. Dellino GI, Cittaro D, Piccioni R, Luzi L, Banfi S, Segalla S, et al. Genome-wide mapping of human DNA-replication origins: Levels of transcription at ORC1 sites regulate origin selection and replication timing. *Genome Res.* 2013;23(1):1–11.
  12. Krude T. Initiation of human DNA replication in vitro using nuclei from cells arrested at an initiation-competent state. *J Biol Chem.* 2000;275(18):13699–707.
  13. Marheineke K, Hyrien O, Krude T. Visualization of bidirectional initiation of chromosomal DNA replication in a human cell free system. *Nucleic Acids Res.* 2005;33(21):6931–41.
  14. Szüts D, Krude T. Cell cycle arrest at the initiation step of human chromosomal DNA replication causes DNA damage. *J Cell Sci.* 2004;117(21):4897–908.

15. Szüts D, Christov C, Kitching L, Krude T. Distinct populations of human PCNA are required for initiation of chromosomal DNA replication and concurrent DNA repair. *Exp Cell Res.* 2005;311(2):240–50.
16. Rhodes D, Lipps HJ. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res.* 2015;43(18):8627–37.
17. Piovesan A, Pelleri MC, Antonaros F, Strippoli P, Caracausi M, Vitale L. On the length, weight and GC content of the human genome. *BMC Res Notes* [Internet]. 2019;12(1):1–7. Available from: <https://doi.org/10.1186/s13104-019-4137-z>
18. Gardiner-Garden M, Frommer M. CpG Islands in Vertebrate Genomes. *J Mol Biol.* 1987;196:261–82.
19. Valton AL, Prioleau MN. G-Quadruplexes in DNA Replication: A Problem or a Necessity? *Trends Genet* [Internet]. 2016;32(11):697–706. Available from: <http://dx.doi.org/10.1016/j.tig.2016.09.004>
20. Foulk MS, Urban JM, Casella C, Gerbi SA. Characterizing and controlling intrinsic biases of lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-quadruplex motifs around a subset of human replication origins. *Genome Res.* 2015;125(5):725–35.
21. Cayrou C, Coulombe P, Puy A, Rialle S, Kaplan N, Segal E, et al. New insights into replication origin characteristics in metazoans. *Cell Cycle.* 2012;11(4):658–67.
22. Pherson M, Misulovin Z, Gause M, Dorsett D. Cohesin occupancy and composition at enhancers and promoters are linked to DNA replication origin proximity in *Drosophila*. *Genome Res.* 2019;29(4):602–12.
23. Sequeira-Mendes J, Díaz-Uriarte R, Apeaile A, Huntley D, Brockdorff N, Gómez M. Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genet.* 2009;5(4).
24. Delgado S, Gómez M, Bird A, Antequera F. Initiation of DNA replication at CpG islands in mammalian chromosomes. *EMBO J.* 1998;17(8):2426–35.
25. Prioleau MN. CpG Islands: Starting blocks for replication and transcription. *PLoS Genet.* 2009;5(4):4–5.
26. Long HK, King HW, Patient RK, Odom DT, Klose RJ. Protection of CpG islands from DNA methylation is DNA-encoded and evolutionarily conserved. *Nucleic Acids Res.* 2016;44(14):6693–706.
27. Prioleau MN, MacAlpine DM. DNA replication origins—Where do we begin? *Genes*

- Dev. 2016;30(15):1683–97.
28. Marks AB, Fu H, Aladjem MI. Regulation of replication origins. *Adv Exp Med Biol.* 2017;1042:43–59.
  29. Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, et al. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature.* 2010;464(7291):1082–6.
  30. Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr ARW, James KD, Turner DJ, et al. Orphan CpG Islands Identify numerous conserved promoters in the mammalian genome. *PLoS Genet.* 2010;6(9).
  31. Li H, Liefke R, Jiang J, Kurland JV, Tian W, Deng P, et al. Polycomb-like proteins link the PRC2 complex to CpG islands. *Nature.* 2017;549(7671):287–91.
  32. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, et al. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell.* 2006;125(2):315–26.
  33. Sen S, Block KF, Pasini A, Baylin SB, Easwaran H. Genome-wide positioning of bivalent mononucleosomes. *BMC Med Genomics [Internet].* 2016;9(1):1–14. Available from: <http://dx.doi.org/10.1186/s12920-016-0221-6>
  34. Smith OK, Kim R, Fu H, Martin MM, Lin CM, Utani K, et al. Distinct epigenetic features of differentiation-regulated replication origins. *Epigenetics and Chromatin.* 2016;9(1):1–17.
  35. Cadoret JC, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, Duret L, et al. Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc Natl Acad Sci U S A.* 2008;105(41):15837–42.
  36. Delgado S, Gómez M, Bird A, Antequera F. Initiation of DNA replication at CpG islands in mammalian chromosomes. *EMBO J.* 1998;17(8):2426–35.
  37. Valton AL, Hassan-Zadeh V, Lema I, Boggetto N, Alberti P, Saintomé C, et al. G4 motifs affect origin positioning and efficiency in two vertebrate replicators. *EMBO J.* 2014;33(7):732–46.
  38. Martin MM, Ryan M, Kim RG, Zakas AL, Fu H, Lin CM, et al. Genome-wide depletion of replication initiation events in highly transcribed regions. *Genome Res.* 2011;21(11):1822–32.
  39. Chen Y-H, Keegan S, Kahli M, Tonzi P, Fenyo D, Huang TT, et al. Transcription shapes DNA replication initiation and termination in human cells. *Nat Struct Mol Biol.*

- 2019;26(1):67–77.
40. Aladjem MI. Replication in context: Dynamic regulation of DNA replication patterns in metazoans. *Nat Rev Genet.* 2007;8(8):588–600.
  41. Blackledge NP, Klose RJ. CpG island chromatin: A platform for gene regulation. *Epigenetics.* 2011;6(2):147–52.
  42. Liu Y, Bondarenko V, Ninfa A, Studitsky VM. DNA supercoiling allows enhancer action over a large distance. *Proc Natl Acad Sci U S A.* 2001;98(26):14883–8.
  43. Kuo AJ, Song J, Cheung P, Ishibe-Murakami S, Yamazoe S, Chen JK, et al. ORC1 BAH domain links H4K20 to DNA replication licensing and Meier-Gorlin syndrome. *Nature* [Internet]. 2012;484(7392):115–9. Available from: <http://dx.doi.org/10.1038/nature10956>0Apapers3://publication/doi/10.1038/nature10956
  44. Brustel J, Kirstein N, Izard F, Grimaud C, Prorok P, Cayrou C, et al. Histone H4K20 trimethylation at late-firing origins ensures timely heterochromatin replication. *EMBO J.* 2017;36(18):2726–41.
  45. Oda H, Okamoto I, Murphy N, Chu J, Price SM, Shen MM, et al. Monomethylation of Histone H4-Lysine 20 Is Involved in Chromosome Structure and Stability and Is Essential for Mouse Development. *Mol Cell Biol.* 2009;29(8):2278–95.
  46. Brustel J, Tardat M, Kirsh O, Grimaud C, Julien E. Coupling mitosis to DNA replication: The emerging role of the histone H4-lysine 20 methyltransferase PR-Set7. *Trends Cell Biol* [Internet]. 2011;21(8):452–60. Available from: <http://dx.doi.org/10.1016/j.tcb.2011.04.006>
  47. Schotta G, Sengupta R, Kubicek S, Malin S, Kauer M, Callén E, et al. A chromatin-wide transition to H4K20 monomethylation impairs genome integrity and programmed DNA rearrangements in the mouse. *Genes Dev.* 2008;22(15):2048–61.
  48. Kumagai A, Dunphy WG. Binding of the Treslin-MTBP Complex to Specific Regions of the Human Genome Promotes the Initiation of DNA Replication. *Cell Rep.* 2020;32(12):1–44.
  49. Kumagai A, Dunphy WG. MTBP, the partner of Treslin, contains a novel DNA-binding domain that is essential for proper initiation of DNA replication. *Mol Biol Cell.* 2017;28(22):2998–3012.
  50. Long H, Zhang L, Lv M, Wen Z, Zhang W, Chen X, et al. H2A.Z facilitates licensing and activation of early replication origins. *Nature* [Internet]. 2020;577(7791):576–81.

Available from: <http://dx.doi.org/10.1038/s41586-019-1877-9>

51. Fu H, Maunakea AK, Martin MM, Huang L, Zhang Y, Ryan M, et al. Methylation of Histone H3 on Lysine 79 Associates with a Group of Replication Origins and Helps Limit DNA Replication Once per Cell Cycle. *PLoS Genet.* 2013;9(6):1–14.
52. Wood K, Tellier M, Murphy S. DOT1L and H3K79 methylation in transcription and genomic stability. *Biomolecules.* 2018;8(1):1–16.
53. Bae S, Lesch BJ. H3K4me1 Distribution Predicts Transcription State and Poising at Promoters. *Front Cell Dev Biol.* 2020;8(May):1–11.
54. Local A, Huang H, Albuquerque CP, Singh N, Lee AY, Wang W, et al. Identification of H3K4me1-associated proteins at mammalian enhancers. *Nat Genet [Internet].* 2018;50(1):73–82. Available from: <http://dx.doi.org/10.1038/s41588-017-0015-6>
55. Voigt P, Tee WW, Reinberg D. A double take on bivalent promoters. *Genes Dev.* 2013;27(12):1318–38.
56. Chammas P, Mocavini I, Di Croce L. Engaging chromatin: PRC2 structure meets function. *Br J Cancer [Internet].* 2019;2(September). Available from: <http://dx.doi.org/10.1038/s41416-019-0615-2>
57. Lubelsky Y, Prinz JA, DeNapoli L, Li Y, Belsky JA, MacAlpine DM. DNA replication and transcription programs respond to the same chromatin cues. *Genome Res.* 2014;24(7):1102–14.
58. Bartke T, Vermeulen M, Xhemalce B, Robson SC, Mann M, Kouzarides T. Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell [Internet].* 2010;143(3):470–84. Available from: <http://dx.doi.org/10.1016/j.cell.2010.10.012>
59. Schmitges FW, Prusty AB, Faty M, Stützer A, Lingaraju GM, Aiwazian J, et al. Histone Methylation by PRC2 Is Inhibited by Active Chromatin Marks. *Mol Cell.* 2011;42(3):330–41.
60. Yuan W, Xu M, Huang C, Liu N, Chen S, Zhu B. H3K36 methylation antagonizes PRC2-mediated H3K27 methylation. *J Biol Chem.* 2011;286(10):7983–9.
61. Rizzardi LF, Dorn ES, Strahl BD, Cook JG. DNA replication origin function is Promoted by H3K4 di-methylation in *Saccharomyces cerevisiae*. *Genetics.* 2012;192(2):371–84.
62. Pekowska A, Benoukraf T, Ferrier P, Spicuglia S. A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Res.* 2010;20(11):1493–502.

63. Barski A, Cuddapah S, Cui K, Roh TY, Schonnes DE, Wang Z, et al. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*. 2007;129(4):823–37.
64. Koch CM, Andrews RM, Flicek P, Dillon SC, Karaöz U, Clelland GK, et al. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res*. 2007;17(6):691–707.
65. Bellush JM, Whitehouse I. DNA replication through a chromatin environment. *Philos Trans R Soc B Biol Sci*. 2017;372(1731).
66. Turner BM. Reading signals on the nucleosome with a new nomenclature for modified histones. *Nat Struct Mol Biol*. 2005;12(2):110–2.
67. Ghare SS, Joshi-Barve S, Moghe A, Patil M, Barker DF, Gobejishvili L, et al. Coordinated histone H3 methylation and acetylation regulates physiologic and pathologic Fas Ligand gene expression in human CD4+ T cells. *J Immunol*. 2014;193(1):412–21.
68. Kylie K, Romero J, Lindamulage IKS, Knockleby J, Lee H. Dynamic regulation of histone H3K9 is linked to the switch between replication and transcription at the Dbf4 origin-promoter locus. *Cell Cycle [Internet]*. 2016;15(17):2321–35. Available from: <http://dx.doi.org/10.1080/15384101.2016.1201254>
69. Du Q, Bert SA, Armstrong NJ, Caldon CE, Song JZ, Nair SS, et al. Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. *Nat Commun [Internet]*. 2019;10(1):1–15. Available from: <http://dx.doi.org/10.1038/s41467-019-08302-1>
70. Julienne H, Zoufir A, Audit B, Arneodo A. Human Genome Replication Proceeds through Four Chromatin States. *PLoS Comput Biol*. 2013;9(10).
71. Creighton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A*. 2010;107(50):21931–6.
72. Beck DB, Burton A, Oda H, Ziegler-Birling C, Torres-Padilla ME, Reinberg D. The role of PR-Set7 in replication licensing depends on Suv4-20h. *Genes Dev*. 2012;26(23):2580–9.
73. Tardat M, Brustel J, Kirsh O, Lefevbre C, Callanan M, Sardet C, et al. The histone H4 Lys 20 methyltransferase PR-Set7 regulates replication origins in mammalian cells. *Nat Cell Biol*. 2010;12(11):1086–93.
74. Fragkos M, Ganier O, Coulombe P, Méchali M. DNA replication origin activation in

- space and time. *Nat Rev Mol Cell Biol.* 2015;16(6):360–74.
75. SeqMonk. Babraham Bioinformatics Seqmonk Project [Internet]. Vol. Version 1. [cited 2020 Sep 20]. Available from:  
<http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/>
  76. Hughes AL, Kelley JR, Klose RJ. Understanding the interplay between CpG island-associated gene promoters and H3K4 methylation. *Biochim Biophys Acta - Gene Regul Mech* [Internet]. 2020;1863(8):194567. Available from:  
<https://doi.org/10.1016/j.bbagr.2020.194567>
  77. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. *Cell.* 2007;130(1):77–88.
  78. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature.* 2007;448(7153):553–60.
  79. Rose NR, Klose RJ. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim Biophys Acta - Gene Regul Mech* [Internet]. 2014;1839(12):1362–72. Available from:  
<http://dx.doi.org/10.1016/j.bbagr.2014.02.007>
  80. Shin Voo K, Carlone DL, Jacobsen BM, Flodin A, Skalnik DG. Cloning of a Mammalian Transcriptional Activator That Binds Unmethylated CpG Motifs and Shares a CXXC Domain with DNA Methyltransferase, Human Trithorax, and Methyl-CpG Binding Domain Protein 1. *Mol Cell Biol.* 2000;20(6):2108–21.
  81. Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes Dev.* 2011;25:1010–22.

*Chapter 6: The role of hY RNAs and xNuRD for the activation of human DNA replication origins, as determined by ds-iniSeq*

1. Christov CP, Gardiner TJ, Szuts D, Krude T. Functional Requirement of Noncoding Y RNAs for Human Chromosomal DNA Replication. *Mol Cell Biol.* 2006;26(18):6993–7004.
2. Christov CP, Dingwell KS, Skehel M, Wilkes HS, Sale JE, Smith JC, et al. A NuRD Complex from *Xenopus laevis* Eggs Is Essential for DNA Replication during Early

- Embryogenesis. *Cell Rep* [Internet]. 2018;22(9):2265–78. Available from: <https://doi.org/10.1016/j.celrep.2018.02.015>
3. Lemaitre J-M, Geraud G, Mechali M. Dynamics of the Genome during Early *Xenopus laevis* Development: Karyomeres As Independent Units of Replication. *J Cell Biol.* 1998;142(5):1159–66.
  4. Collart C, Christov CP, Smith JC, Krude T. The Midblastula Transition Defines the Onset of Y RNA-Dependent DNA Replication in *Xenopus laevis*. *Mol Cell Biol.* 2011;31(18):3857–70.
  5. Mahbubani HM, Paull T, Eider JK, Blow JJ. DNA replication initiates at multiple sites on plasmid DNA in *xenopus* egg extracts. *Nucleic Acids Res.* 1992;20(7):1457–62.
  6. Blow JJ. Control of chromosomal DNA replication in the early *Xenopus* embryo. *EMBO J.* 2001;20(13):3293–7.
  7. Rivera-Mulia JC, Gilbert DM. Replication timing and transcriptional control: beyond cause and effect—part III. *Curr Opin Cell Biol.* 2016;40:168–78.
  8. Dellino GI, Cittaro D, Piccioni R, Luzi L, Banfi S, Segalla S, et al. Genome-wide mapping of human DNA-replication origins: Levels of transcription at ORC1 sites regulate origin selection and replication timing. *Genome Res.* 2013;23(1):1–11.
  9. Christov CP, Trivier E, Krude T. Noncoding human Y RNAs are overexpressed in tumours and required for cell proliferation. *Br J Cancer.* 2008;98(5):981–8.
  10. Gardiner TJ, Christov CP, Langley AR, Krude T. A conserved motif of vertebrate Y RNAs essential for chromosomal DNA replication. *Rna.* 2009;15(7):1375–85.
  11. Krude T, Christov CP, Hyrien O, Marheineke K. Y RNA functions at the initiation step of mammalian chromosomal DNA replication. *J Cell Sci.* 2009;122(16):2836–45.
  12. Denslow SA, Wade PA. The human Mi-2/NuRD complex and gene regulation. *Oncogene.* 2007;26(37):5433–8.
  13. Lai AY, Wade PA. NuRD: A multi-faceted chromatin remodeling complex in regulating cancer biology. *Nat Rev Cancer.* 2011;11(8):588–96.
  14. Hoffmeister H, Fuchs A, Erdel F, Pinz S, Gröbner-Ferreira R, Bruckmann A, et al. CHD3 and CHD4 form distinct NuRD complexes with different yet overlapping functionality. *Nucleic Acids Res.* 2017;45(18):10534–54.
  15. Millard CJ, Varma N, Saleh A, Morris K, Watson PJ, Bottrill AR, et al. The structure of the core NuRD repression complex provides insights into its interaction with

- chromatin. *Elife*. 2016;5(APRIL2016):1–21.
16. Le Guezennec X, Vermeulen M, Brinkman AB, Hoeijmakers WAM, Cohen A, Lasonder E, et al. MBD2/NuRD and MBD3/NuRD, Two Distinct Complexes with Different Biochemical and Functional Properties. *Mol Cell Biol*. 2006;26(3):843–51.
  17. Günther K, Rust M, Leers J, Boettger T, Scharfe M, Jarek M, et al. Differential roles for MBD2 and MBD3 at methylated CpG islands, active promoters and binding to exon sequences. *Nucleic Acids Res*. 2013;41(5):3010–21.
  18. Shima N, Pederson KD. Dormant origins as a built-in safeguard in eukaryotic DNA replication against genome instability and disease development Naoko. *DNA Repair (Amst)*. 2017;56:166–73.
  19. Blow JJ, Quan Ge X, Jackson DA, Blow J. How dormant origins promote complete genome replication. *Trends Biochem Sci*. 2011;36(8):405–14.
  20. Mcintosh D, Blow JJ, Jackson D, Wang X, Rudner DZ, Mcintosh D, et al. Dormant Origins, the Licensing Checkpoint, and the Response to Replicative Stresses. *Cold Spring Harb Perspect Biol*. 2012;4(a012955):1–10.
  21. Sugimoto N, Maehara K, Yoshida K, Ohkawa Y, Fujita M. Genome-wide analysis of the spatiotemporal regulation of firing and dormant replication origins in human cells. *Nucleic Acids Res*. 2018;46(13):6683–96.
  22. Zhang AT, Langley AR, Christov CP, Kheir E, Shafee T, Gardiner TJ, et al. Dynamic interaction of Y RNAs with chromatin and initiation proteins during human DNA replication. *J Cell Sci*. 2011;124(12):2058–69.
  23. Kheir E, Krude T. Non-coding Y RNAs associate with early replicating euchromatin in concordance with the origin recognition complex. *J Cell Sci*. 2017;130(7):1239–50.
  24. Tatsumi Y, Ohta S, Kimura H, Tsurimoto T, Obuse C. The ORC1 cycle in human cells: I. Cell cycle-regulated oscillation of human ORC1. *J Biol Chem*. 2003;278(42):41528–34.
  25. Tsakraklides V, Bell SP. Dynamics of pre-replicative complex assembly. *J Biol Chem*. 2010;285(13):9437–43.
  26. Bell SP, Dutta A. DNA Replication in Eukaryotic Cells. *Annu Rev Biochem*. 2002;71(1):333–74.
  27. Roberts K. The methyltransferase PRMT1 is essential for chromosomal DNA replication in vitro. University of Cambridge; 2017.

28. Bedford MT, Richard S. Arginine methylation: An emerging regulator of protein function. *Mol Cell*. 2005;18(3):263–72.
29. Tang J, Frankel A, Cook RJ, Kim S, Paik WK, Williams KR, et al. PRMT1 is the predominant type I protein arginine methyltransferase in mammalian cells. *J Biol Chem*. 2000;275(11):7723–30.
30. Nicholson TB, Chen T, Richard S. The physiological and pathophysiological role of PRMT1-mediated protein arginine methylation. *Pharmacol Res*. 2009;60(6):466–74.
31. Strahl BD, Briggs SD, Brame CJ, Caldwell JA, Koh SS, Ma H, et al. Methylation of histone H4 at arginine 3 occurs in vivo and is mediated by the nuclear receptor coactivator PRMT1. *Curr Biol*. 2001;11(12):996–1000.
32. Bedford MT, Clarke SG. Protein Arginine Methylation in Mammals: Who, What, and Why. *Mol Cell*. 2009;33(1):1–13.
33. Cha B, Jho EH. Protein arginine methyltransferases (PRMTs) as therapeutic targets. *Expert Opin Ther Targets*. 2012;16(7):651–64.
34. Beacon TH, Xu W, Davie JR. Genomic landscape of transcriptionally active histone arginine methylation marks, H3R2me2s and H4R3me2a, relative to nucleosome depleted regions. *Gene [Internet]*. 2020;742(March):144593. Available from: <https://doi.org/10.1016/j.gene.2020.144593>
35. Basta J, Rauchman M. The Nucleosome Remodeling and Deacetylase Complex in Development and Disease. *Transl Epigenetics to Clin*. 2017;165(1):37–72.

*Chapter 7: The impact of Y RNAs and xNuRD on replication elongation using density-substitution elongation-site sequencing (ds-eloSeq)*

1. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):1–9.
2. Szybalski W. [124] Use of cesium sulfate for equilibrium density gradient centrifugation. In: *Methods in Enzymology [Internet]*. 1968. p. 330–60. Available from: [https://doi.org/10.1016/0076-6879\(67\)12149-6](https://doi.org/10.1016/0076-6879(67)12149-6)
3. Christov CP, Dingwell KS, Skehel M, Wilkes HS, Sale JE, Smith JC, et al. A NuRD

Complex from *Xenopus laevis* Eggs Is Essential for DNA Replication during Early Embryogenesis. *Cell Rep* [Internet]. 2018;22(9):2265–78. Available from: <https://doi.org/10.1016/j.celrep.2018.02.015>

4. SeqMonk. Babraham Bioinformatics Seqmonk Project [Internet]. Vol. Version 1. [cited 2020 Sep 20]. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/>
5. Krude T, Christov CP, Hyrien O, Marheineke K. Y RNA functions at the initiation step of mammalian chromosomal DNA replication. *J Cell Sci*. 2009;122(16):2836–45.
6. Zhang AT, Langley AR, Christov CP, Kheir E, Shafee T, Gardiner TJ, et al. Dynamic interaction of Y RNAs with chromatin and initiation proteins during human DNA replication. *J Cell Sci*. 2011;124(12):2058–69.
7. Zhong Y, Nellimooti T, Peace JM, Knott SRV, Villwock SK, Yee JM, et al. The level of origin firing inversely affects the rate of replication fork progression. *J Cell Biol*. 2013;201(3):373–83.
8. Dellino GI, Cittaro D, Piccioni R, Luzi L, Banfi S, Segalla S, et al. Genome-wide mapping of human DNA-replication origins: Levels of transcription at ORC1 sites regulate origin selection and replication timing. *Genome Res*. 2013;23(1):1–11.

*Chapter 8: A role for the chromatin remodelling complex, Polycomb Repressive Complex 2 (PRC2), in human replication initiation*

1. Christov CP, Gardiner TJ, Szuts D, Krude T. Functional Requirement of Noncoding Y RNAs for Human Chromosomal DNA Replication. *Mol Cell Biol*. 2006;26(18):6993–7004.
2. Balcerak A, Trebinska-Stryjewska A, Konopinski R, Wakula M, Grzybowska EA. RNA-protein interactions: Disorder, moonlighting and junk contribute to eukaryotic complexity. *Open Biol*. 2019;9(6).
3. Kowalski M. Investigating the function of non-coding RNAs in chromosomal DNA replication. University of Cambridge. 2014.
4. van Mierlo G, Veenstra GJC, Vermeulen M, Marks H. The Complexity of PRC2 Subcomplexes. *Trends Cell Biol* [Internet]. 2019;29(8):660–71. Available from: <https://doi.org/10.1016/j.tcb.2019.05.004>

5. Chammas P, Mocavini I, Di Croce L. Engaging chromatin: PRC2 structure meets function. *Br J Cancer* [Internet]. 2019;2(September). Available from: <http://dx.doi.org/10.1038/s41416-019-0615-2>
6. Shi Y, Wang XX, Zhuang YW, Jiang Y, Melcher K, Xu HE. Structure of the PRC2 complex and application to drug discovery. *Acta Pharmacol Sin*. 2017;38(7):963–76.
7. Son J, Shen SS, Margueron R, Reinberg D. Nucleosome-binding activities within JARID2 and EZH1 regulate the function of PRC2 on chromatin. *Genes Dev*. 2013;27(24):2663–77.
8. Margueron R, Reinberg D. The Polycomb Complex PRC2 and its Mark in Life. *Nature*. 2011;469(7330):343–9.
9. Schuettengruber B, Cavalli G. Recruitment of Polycomb group complexes and their role in the dynamic regulation of cell fate choice. *Development*. 2009;136(21):3531–42.
10. Simon JA, Kingston RE. Mechanisms of Polycomb gene silencing: Knowns and unknowns. *Nat Rev Mol Cell Biol* [Internet]. 2009;10(10):697–708. Available from: <http://dx.doi.org/10.1038/nrm2763>
11. Kasinath V, Faini M, Poepsel S, Reif D, Feng XA, Stjepanovic G, et al. Structures of human PRC2 with its cofactors AEBP2 and JARID2. *Science* (80- ). 2018;359(6378):940–4.
12. McGinty RK, Henrici RC, Tan S. Crystal structure of the PRC1 ubiquitylation module bound to the nucleosome. *Nature*. 2014;514(7524):591–6.
13. Gao Z, Zhang J, Bonasio R, Strino F, Sawai A, Parisi F, et al. PCGF Homologs, CBX Proteins, and RYBP Define Functionally Distinct PRC1 Family Complexes. *Mol Cell*. 2012;45(3):344–56.
14. Francis NJ, Follmer NE, Simon MD, Aghia G, Butler JD. Polycomb proteins remain bound to chromatin and DNA during DNA replication in vitro. *Cell* [Internet]. 2009;137(1):110–22. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf>
15. Tan JZ, Yan Y, Wang XX, Jiang Y, Xu HE. EZH2: Biology, disease, and structure-based drug discovery. *Acta Pharmacol Sin*. 2014;35(2):161–74.
16. Czermin B, Melfi R, McCabe D, Seitz V, Imhof A, Pirrotta V. Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. *Cell*. 2002;111(2):185–96.

17. Lee C-H, Holder M, Grau D, Saldaña-Meyer R, Yu J-R, Ganai RA, et al. Distinct stimulatory mechanisms regulate the catalytic activity of Polycomb Repressive Complex 2 (PRC2). *Mol Cell*. 2018;70(3):435–48.
18. Abel KJ, Brody LC, Valdes JM, Erdos MR, McKinley DR, Castilla LH, et al. Characterization of EZH1, a human homolog of Drosophila enhancer of zeste near BRCA1. *Genomics*. 1996;37(2):161–71.
19. Margueron R, Li G, Sarma K, Blais A, Zavadil J, Woodcock CL, et al. Ezh1 and Ezh2 Maintain Repressive Chromatin through Different Mechanisms. *Mol Cell*. 2008;32(4):503–18.
20. Yu JR, Lee CH, Oksuz O, Stafford JM, Reinberg D. PRC2 is high maintenance. *Genes Dev*. 2019;33(15–16):903–35.
21. Margueron R, Justin N, Ohno K, Sharpe ML, Son J, Iijima WJD, et al. Role of the polycomb protein Eed in the propagation of repressive histone marks. *Nature*. 2009;461(7265):762–7.
22. Cao R, Zhang Y. SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex. *Mol Cell*. 2004;15(1):57–67.
23. Pasini D, Bracken AP, Jensen MR, Denchi EL, Helin K. Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. *EMBO J*. 2004;23(20):4061–71.
24. Wu H, Zeng H, Dong A, Li F, He H, Senisterra G, 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):1–12.
25. Bratkowski M, Yang X, Liu X. Polycomb repressive complex 2 in an autoinhibited state. *J Biol Chem*. 2017;292(32):13323–32.
26. Antonysamy S, Condon B, Druzina Z, Bonanno JB, Gheyi T, Zhang F, et al. Structural context of disease-associated mutations and putative mechanism of autoinhibition revealed by X-Ray crystallographic analysis of the EZH2-SET domain. *PLoS One*. 2013;8(12):1–15.
27. Lee C-H, Yu J-R, Kumar S, Jin Y, LeRoy G, Bhanu N, et al. Allosteric activation dictates PRC2 activity independent of its recruitment to chromatin. *Mol Cell*. 2018;70(3):422–34.
28. Cao Q, Wang X, Zhao M, Yang R, Malik R, Qiao Y, et al. The Central Role of EED in the Orchestration of Polycomb Group Complexes. *Nat Commun*. 2014;5(3127):1–26.

29. Han Z, Xing X, Hu M, Zhang Y, Liu P, Chai J. Structural Basis of EZH2 Recognition by EED. *Structure*. 2007;15(10):1306–15.
30. Montgomery ND, Yee D, Montgomery SA, Magnuson T. Molecular and functional mapping of EED motifs required for PRC2-dependent histone methylation. *J Mol Biol*. 2007;374(5):1145–57.
31. Chen S, Jiao L, Shubbar M, Yang X, Liu X. Unique Structural Platforms of Suz12 Dictate Distinct Classes of PRC2 for Chromatin Binding. *Mol Cell* [Internet]. 2018;69(5):840-852.e5. Available from: <https://doi.org/10.1016/j.molcel.2018.01.039>
32. Youmans DT, Schmidt JC, Cech TR. Live-cell imaging reveals the dynamics of PRC2 and recruitment to chromatin by SUZ12-associated subunits. *Genes Dev*. 2018;32(11–12):794–805.
33. Ketel CS, Andersen EF, Vargas ML, Suh J, Strome S, Simon JA. Subunit Contributions to Histone Methyltransferase Activities of Fly and Worm Polycomb Group Complexes. *Mol Cell Biol*. 2005;25(16):6857–68.
34. Schmitges FW, Prusty AB, Faty M, Stützer A, Lingaraju GM, Aiwazian J, et al. Histone Methylation by PRC2 Is Inhibited by Active Chromatin Marks. *Mol Cell*. 2011;42(3):330–41.
35. Ciferri C, Lander GC, Maiolica A, Herzog F, Aebersold R, Nogales E. Molecular architecture of human polycomb repressive complex 2. *Elife*. 2012;2012(1):1–22.
36. Wassef M, Luscan A, Aflaki S, Zielinski D, Jansen PWTC, Baymaz HI, et al. EZH1/2 function mostly within canonical PRC2 and exhibit proliferation-dependent redundancy that shapes mutational signatures in cancer. *Proc Natl Acad Sci U S A*. 2019;116(13):6075–80.
37. Kanhere A, Viiri K, Araújo CC, Rasaiyaah J, Bouwman RD, Whyte WA, et al. Short RNAs are transcribed from repressed Polycomb target genes and interact with Polycomb Repressive Complex-2. 2010;38(5):675–88.
38. Beltran M, Yates CM, Skalska L, Dawson M, Reis FP, Viiri K, et al. The interaction of PRC2 with RNA or chromatin is mutually antagonistic. *Genome Res*. 2016;26(7):896–907.
39. Murzina N V., Pei XY, Zhang W, Sparkes M, Vicente-Garcia J, Pratap JV, et al. Structural Basis for the Recognition of Histone H4 by the Histone-Chaperone RbAp46. *Structure* [Internet]. 2008;16(7):1077–85. Available from: <http://dx.doi.org/10.1016/j.str.2008.05.006>

40. Nekrasov M, Wild B, Müller J. Nucleosome binding and histone methyltransferase activity of *Drosophila* PRC2. *EMBO Rep.* 2005;6(4):348–53.
41. Abbey M, Trush V, Gibson E, Vedadi M. Targeting Human Retinoblastoma Binding Protein 4 (RBBP4) and 7 (RBBP7). *bioRxiv - Prepr [Internet]*. 2018;Submitted. Available from: <https://www.biorxiv.org/content/early/2018/04/18/303537>
42. Zhang Y, Ng HH, Erdjument-Bromage H, Tempst P, Bird A, Reinberg D. Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev.* 1999;13(15):1924–35.
43. Lai AY, Wade PA. NuRD: A multi-faceted chromatin remodeling complex in regulating cancer biology. *Nat Rev Cancer.* 2011;11(8):588–96.
44. Yang X-J, Seto E. The Rpd3/Hda1 family of lysine deacetylases: from bacteria and yeast to mice and men. *Nat Rev Genet.* 2008;9(3):206–18.
45. Loyola A, Almouzni G. Histone chaperones, a supporting role in the limelight. *Biochim Biophys Acta - Gene Struct Expr.* 2004;1677(1–3):3–11.
46. Xu C, Min J. Structure and function of WD40 domain proteins. *Protein Cell.* 2011;2(3):202–14.
47. Christov CP, Dingwell KS, Skehel M, Wilkes HS, Sale JE, Smith JC, et al. A NuRD Complex from *Xenopus laevis* Eggs Is Essential for DNA Replication during Early Embryogenesis. *Cell Rep [Internet]*. 2018;22(9):2265–78. Available from: <https://doi.org/10.1016/j.celrep.2018.02.015>
48. Justin N, Zhang Y, Tarricone C, Martin SR, Chen S, Underwood E, et al. Structural basis of oncogenic histone H3K27M inhibition of human polycomb repressive complex 2. *Nat Commun [Internet]*. 2016;7:11316. Available from: <http://dx.doi.org/10.1038/ncomms11316>
49. Sanulli S, Justin N, Teissandier A, Ancelin K, Portoso M, Caron M, et al. Jarid2 Methylation via the PRC2 Complex Regulates H3K27me3 Deposition during Cell Differentiation. *Mol Cell.* 2015;57(5):769–83.
50. Tokodai Y, Yakushiji F. Polycomb Repressive Complex 2: Modulator Development for Functional Regulation of a Multiprotein Complex by Using Structural Information. *ChemBioChem.* 2019;20(16):2046–53.
51. Chittock EC, Latwiel S, Miller TCR, Müller CW. Molecular architecture of polycomb repressive complexes. *Biochem Soc Trans.* 2017;45(1):193–205.
52. McCabe MT, Ott HM, Ganji G, Korenchuk S, Thompson C, Van Aller GS, et al. EZH2

- inhibition as a therapeutic strategy for lymphoma with EZH2-activating mutations. *Nature* [Internet]. 2012;492(7427):108–12. Available from: <http://dx.doi.org/10.1038/nature11606>
53. Kim KH, Roberts CWM. Targeting EZH2 in cancer. *Nat Med*. 2016;22(2):128–34.
  54. Greer EL, Shi Y. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat Rev Genet*. 2012;13(5):343–57.
  55. Xu B, Konze KD, Jin J, Wang GG. Targeting EZH2 and PRC2 dependence as novel anticancer therapy. *Exp Hematol*. 2015;43(8):698–712.
  56. Wagner EJ, Carpenter PB. Understanding the language of Lys36 methylation at histone H3. *Nat Rev Mol Cell Biol*. 2012;13(2):115–26.
  57. Yan J, Ng SB, Tay JLS, Lin B, Koh TL, Tan J, et al. EZH2 overexpression in natural killer/T-cell lymphoma confers growth advantage independently of histone methyltransferase activity. *Blood*. 2013;121(22):4512–20.
  58. Cavalli G. EZH2 goes solo. *Science* (80- ). 2012;338(6113):1430–1.
  59. Kleer CG, Cao Q, Varambally S, Shen R, Ota I, Tomlins SA, et al. EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proc Natl Acad Sci U S A*. 2003;100(20):11606–11.
  60. Varambally S, Dhanasekaran SM, Zhou M, Barrette TR, Kumar-Sinha C, Sanda MG, et al. The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature*. 2002;419(6907):624–9.
  61. Takawa M, Masuda K, Kunizaki M, Daigo Y, Takagi K, Iwai Y, et al. Validation of the histone methyltransferase EZH2 as a therapeutic target for various types of human cancer and as a prognostic marker. *Cancer Sci*. 2011;102(7):1298–305.
  62. Vilorio-Marqués L, Martín V, Diez-Tascón C, González-Sevilla MF, Fernández-Villa T, Honrado E, et al. The role of EZH2 in overall survival of colorectal cancer: A meta-Analysis. *Sci Rep*. 2017;7(1):1–8.
  63. van Vlerken LE, Kiefer CM, Morehouse C, Li Y, Groves C, Wilson SD, et al. EZH2 Is Required for Breast and Pancreatic Cancer Stem Cell Maintenance and Can Be Used as a Functional Cancer Stem Cell Reporter. *Stem Cells Transl Med*. 2013;2(1):43–52.
  64. Ramaglia M, D'Angelo V, Iannotta A, Pinto D, Pota E, Affinita MC, et al. High EZH2 expression is correlated to metastatic disease in pediatric soft tissue sarcomas. *Cancer Cell Int*. 2016;16(1):1–9.

65. Chen Z, Yang P, Li W, He F, Wei J, Zhang T, et al. Expression of EZH2 is associated with poor outcome in colorectal cancer. *Oncol Lett*. 2018;15(3):2953–61.
66. Baugé C, Bazille C, Girard N, Lhuissier E, Boumediene K. Histone methylases as novel drug targets: Developing inhibitors of EZH2. *Future Med Chem*. 2014;6(17):1943–65.
67. Deevy O, Bracken AP. PRC2 functions in development and congenital disorders. *Development*. 2019;146(19):1–13.
68. Gunawan M, Venkatesan N, Loh JT, Wong JF, Berger H, Neo WH, et al. The methyltransferase Ezh2 controls cell adhesion and migration through direct methylation of the extranuclear regulatory protein talin. *Nat Immunol*. 2015;16(5):505–16.
69. Schimmelmann M von, Feinberg PA, Sullivan JM, Ku SM, Badimon A, Duff MK, et al. Polycomb repressive complex 2 (PRC2) silences genes responsible for neurodegeneration. *Nat Neurosci* [Internet]. 2016;19(10):1321–30. Available from: [file:///C:/Users/Carla Carolina/Desktop/Artigos para acrescentar na qualificação/The impact of birth weight on cardiovascular disease risk in the.pdf](file:///C:/Users/Carla%20Carolina/Desktop/Artigos%20para%20acrescentar%20na%20qualifica%C3%A7%C3%A3o/The%20impact%20of%20birth%20weight%20on%20cardiovascular%20disease%20risk%20in%20the.pdf)
70. Bantug GR, Hess C. Glycolysis and EZH2 boost T cell weaponry against tumors. *Nat Immunol*. 2016;17(1):41–2.
71. Deb G, Singh AK, Gupta S. EZH2: Not EZHY (Easy) to Deal. *Mol Cancer Res* [Internet]. 2014;12(5):639–53. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf>
72. Jiang T, Wang Y, Zhou F, Gao G, Ren S, Zhou C. Prognostic value of high EZH2 expression in patients with different types of cancer: A systematic review with meta-analysis. *Oncotarget*. 2016;7(4):4584–97.
73. Hauri S, Comoglio F, Seimiya M, Gerstung M, Glatter T, Hansen K, et al. A High-Density Map for Navigating the Human Polycomb Complexome. *Cell Rep* [Internet]. 2016;17(2):583–95. Available from: <http://dx.doi.org/10.1016/j.celrep.2016.08.096>
74. Li H, Liefke R, Jiang J, Kurland JV, Tian W, Deng P, et al. Polycomb-like proteins link the PRC2 complex to CpG islands. *Nature*. 2017;549(7671):287–91.
75. Perino M, Van Mierlo G, Karemaker ID, Van Genesen S, Vermeulen M, Marks H, et al. MTF2 recruits Polycomb Repressive Complex 2 by helical-shape-selective DNA binding. *Nat Genet* [Internet]. 2018;50(7):1002–10. Available from: <http://dx.doi.org/10.1038/s41588-018-0134-8>

76. Casanova M, Preissner T, Cerase A, Poot R, Yamada D, Li X, et al. Polycomb-like 2 facilitates the recruitment of PRC2 Polycomb group complexes to the inactive X chromosome and to target loci in embryonic stem cells. *Development*. 2011;138(8):1471–82.
77. Choi J, Bachmann AL, Tauscher K, Benda C, Fierz B, Müller J. DNA binding by PHF1 prolongs PRC2 residence time on chromatin and thereby promotes H3K27 methylation. *Nat Struct Mol Biol*. 2017;24(12):1039–47.
78. Hunkapiller J, Shen Y, Diaz A, Cagney G, McCleary D, Ramalho-Santos M, et al. Polycomb-like 3 promotes polycomb repressive complex 2 binding to CpG islands and embryonic stem cell self-renewal. *PLoS Genet*. 2012;8(3).
79. Beringer M, Pisano P, Di Carlo V, Blanco E, Chammas P, Vizán P, et al. EPOP Functionally Links Elongin and Polycomb in Pluripotent Stem Cells. *Mol Cell*. 2016;64(4):645–58.
80. Smits AH, Jansen PWTC, Poser I, Hyman AA, Vermeulen M. Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. *Nucleic Acids Res*. 2013;41(1):1–8.
81. Shi Y, Ma H lei, Zhuang Y wen, Wang X xi, Jiang Y, Xu HE. C10ORF12 modulates PRC2 histone methyltransferase activity and H3K27me3 levels. *Acta Pharmacol Sin* [Internet]. 2019;40(11):1457–65. Available from: <http://dx.doi.org/10.1038/s41401-019-0247-3>
82. Aso T, Lane WS, Conaway JW, Conaway RC. Elongin (SIII): A multisubunit regulator of elongation by RNA polymerase II. *Science* (80- ). 1995;269(5229):1439–43.
83. Holloch D, Margueron R. Mechanisms Regulating PRC2 Recruitment and Enzymatic Activity. *Trends Biochem Sci* [Internet]. 2017;42(7):531–42. Available from: <http://dx.doi.org/10.1016/j.tibs.2017.04.003>
84. Kalb R, Latwiel S, Baymaz HI, Jansen PWTC, Müller CW, Vermeulen M, et al. Histone H2A monoubiquitination promotes histone H3 methylation in Polycomb repression. *Nat Struct Mol Biol*. 2014;21(6):569–71.
85. Xu C, Bian C, Yang W, Galka M, Ouyang H, Chen C, et al. Binding of different histone marks differentially regulates the activity and specificity of polycomb repressive complex 2 (PRC2). *Proc Natl Acad Sci U S A*. 2010;107(45):19266–71.
86. Bartke T, Vermeulen M, Xhemalce B, Robson SC, Mann M, Kouzarides T. Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell* [Internet]. 2010;143(3):470–84. Available from:

<http://dx.doi.org/10.1016/j.cell.2010.10.012>

87. Yuan W, Xu M, Huang C, Liu N, Chen S, Zhu B. H3K36 methylation antagonizes PRC2-mediated H3K27 methylation. *J Biol Chem*. 2011;286(10):7983–9.
88. Jani KS, Jain SU, Ge EJ, Diehl KL, Lundgren SM, Müller MM, et al. Histone H3 tail binds a unique sensing pocket in EZH2 to activate the PRC2 methyltransferase. *Proc Natl Acad Sci U S A*. 2019;116(17):8295–300.
89. Long Y, Bolanos B, Gong L, Liu W, Goodrich KJ, Yang X, et al. Conserved RNA-binding specificity of polycomb repressive complex 2 is achieved by dispersed amino acid patches in EZH2. *Elife*. 2017;6(1):1–23.
90. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Morales DR, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A*. 2009;106(28):11667–72.
91. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, et al. Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Cell*. 2007;129(7):1311–23.
92. Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science (80- )*. 2008;322(5902):750–6.
93. Davidovich C, Wang X, Cifuentes-rojas C, Goodrich KJ, Gooding AR, Lee JT, et al. Towards a Consensus on the Binding Specificity and Promiscuity of PRC2 for RNA. *Mol Cell*. 2015;57(3):552–8.
94. Kretz M, Meister G. RNA Binding of PRC2: Promiscuous or Well Ordered? *Mol Cell* [Internet]. 2014;55(2):157–8. Available from: <http://dx.doi.org/10.1016/j.molcel.2014.07.002>
95. Cifuentes-rojas C, Hernandez AJ, Sarma K, Lee JT. Regulatory interactions between RNA and Polycomb Repressive Complex 2. *Mol Cell* [Internet]. 2014;55(2):171–85. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf>
96. Davidovich C, Zheng L, Goodrich KJ, Cech TR. Promiscuous RNA binding by Polycomb Repressive Complex 2. *Nat Struct Mol Biol*. 2013;20(11):1250–7.
97. Wang X, Paucek RD, Gooding AR, Brown ZZ, Ge EJ, Muir TW, et al. Molecular analysis of PRC2 recruitment to DNA in chromatin and its inhibition by RNA. *Nat Struct Mol Biol*. 2017;24(12):1028–38.
98. Kaneko S, Bonasio R, Saldaña-Meyer R, Yoshida T, Son J, Nishino K, et al.

- Interactions between JARID2 and Noncoding RNAs Regulate PRC2 Recruitment to Chromatin. *Mol Cell*. 2014;53(2):290–300.
99. Davidovich C, Goodrich KJ, Gooding AR, Cech TR. NAR Breakthrough Article: A dimeric state for PRC2. *Nucleic Acids Res*. 2014;42(14):9236–48.
  100. Yan J, Dutta B, Hee YT, Chng WJ. Towards understanding of PRC2 binding to RNA. *RNA Biol* [Internet]. 2019;16(2):176–84. Available from: <https://doi.org/10.1080/15476286.2019.1565283>
  101. Riising EM, Comet I, Leblanc B, Wu X, Johansen JV, Helin K. Gene silencing triggers polycomb repressive complex 2 recruitment to CpG Islands genome wide. *Mol Cell* [Internet]. 2014;55(3):347–60. Available from: <http://dx.doi.org/10.1016/j.molcel.2014.06.005>
  102. Cooper S, Dienstbier M, Hassan R, Schermelleh L, Sharif J, Blackledge NP, et al. Targeting Polycomb to Pericentric Heterochromatin in Embryonic Stem Cells Reveals a Role for H2AK119u1 in PRC2 Recruitment. *Cell Rep* [Internet]. 2014;7(5):1456–70. Available from: <http://dx.doi.org/10.1016/j.celrep.2014.04.012>
  103. Silva J, Mak W, Zvetkova I, Appanah R, Nesterova TB, Webster Z, et al. Establishment of histone H3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes. *Dev Cell*. 2003;4(4):481–95.
  104. Plath K, Fang J, Mlynarczyk-Evans SK, Cao R, Worringer KA, Wang H, et al. Role of histone H3 lysine 27 methylation in X inactivation. *Science* (80- ). 2003;300(5616):131–5.
  105. Li L, Liu B, Wapinski OL, Tsai M-C, Qu K, Zhang J, et al. Targeted disruption of Hotair leads to homeotic transformation and gene de-repression. *Cell Rep*. 2013;5(1):3–12.
  106. Wang X, Davidovich C. Targeting PRC2: RNA offers new opportunities. *Oncotarget*. 2017;8(64):107346–7.
  107. Wei C, Xiao R, Chen L, Cui H, Zhou Y, Xue Y, et al. RBFox2 Binds Nascent RNA to Globally Regulate Polycomb Complex 2 Targeting in Mammalian Genomes. *Mol Cell*. 2016;62(6):875–89.
  108. Chu C, Zhang QC, Rocha ST da, Flynn RA, Bharadwaj M, Calabrese JM, et al. Systematic discovery of Xist RNA binding proteins. *Cell*. 2015;161(2):404–16.
  109. Xu B, On DM, Ma A, Parton T, Konze KD, Pattenden SG, et al. Selective inhibition of EZH2 and EZH1 enzymatic activity by a small molecule suppresses MLL-rearranged

- leukemia. *Blood*. 2015;125(2):346–57.
110. Konze KD, Ma A, Li F, Baryte-Lovejoy D, Parton T, MacNevin CJ, et al. An Orally Bioavailable Chemical Probe of the Lysine Methyltransferases EZH2 and EZH1. *ACS Chem Biol*. 2013;8(6):1324–34.
  111. Krude T, Jackman M, Pines J, Laskey RA. Cyclin/Cdk-dependent initiation of DNA replication in a human cell-free system. *Cell* [Internet]. 1997;88(1):109–19. Available from: [http://dx.doi.org/10.1016/S0092-8674\(00\)81863-2](http://dx.doi.org/10.1016/S0092-8674(00)81863-2)
  112. Kowalski MP, Krude T. Functional roles of non-coding Y RNAs. *Int J Biochem Cell Biol* [Internet]. 2015;66:20–9. Available from: <http://dx.doi.org/10.1016/j.biocel.2015.07.003>
  113. Roberts K. The methyltransferase PRMT1 is essential for chromosomal DNA replication in vitro. University of Cambridge; 2017.
  114. Bedford MT, Richard S. Arginine methylation: An emerging regulator of protein function. *Mol Cell*. 2005;18(3):263–72.
  115. Tang J, Frankel A, Cook RJ, Kim S, Paik WK, Williams KR, et al. PRMT1 is the predominant type I protein arginine methyltransferase in mammalian cells. *J Biol Chem*. 2000;275(11):7723–30.
  116. Nicholson TB, Chen T, Richard S. The physiological and pathophysiological role of PRMT1-mediated protein arginine methylation. *Pharmacol Res*. 2009;60(6):466–74.
  117. Strahl BD, Briggs SD, Brame CJ, Caldwell JA, Koh SS, Ma H, et al. Methylation of histone H4 at arginine 3 occurs in vivo and is mediated by the nuclear receptor coactivator PRMT1. *Curr Biol*. 2001;11(12):996–1000.
  118. Bedford MT, Clarke SG. Protein Arginine Methylation in Mammals: Who, What, and Why. *Mol Cell*. 2009;33(1):1–13.
  119. Cha B, Jho EH. Protein arginine methyltransferases (PRMTs) as therapeutic targets. *Expert Opin Ther Targets*. 2012;16(7):651–64.
  120. Abe T, Sugimura K, Hosono Y, Takami Y, Akita M, Yoshimura A, et al. The histone chaperone facilitates chromatin transcription (FACT) protein maintains normal replication fork rates. *J Biol Chem*. 2011;286(35):30504–12.
  121. Piquet S, Le Parc F, Bai SK, Chevallier O, Adam S, Polo SE. The Histone Chaperone FACT Coordinates H2A.X-Dependent Signaling and Repair of DNA Damage. *Mol Cell*. 2018;72(5):888-901.e7.
  122. Winkler DD, Luger K. The histone chaperone FACT: Structural insights and

- mechanisms for nucleosome reorganization. *J Biol Chem.* 2011;286(21):18369–74.
123. Basta J, Rauchman M. The Nucleosome Remodeling and Deacetylase Complex in Development and Disease. *Transl Epigenetics to Clin.* 2017;165(1):37–72.
  124. Zhang Q, Mckenzie NJ, Warneford-Thomson R, Mckenzie NJ, Gail EH, Flanigan SF, et al. RNA exploits an exposed regulatory site to inhibit the enzymatic activity of PRC2. *Nat Struct Mol Biol [Internet].* 2019;26(8):237–47. Available from: <https://doi.org/10.1038/s41594-019-0197-y>
  125. Langley AR, Gräf S, Smith JC, Krude T. Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). *Nucleic Acids Res.* 2016;44(21):10230–47.
  126. Mesner LD, Valsakumar V, Cieslik M, Pickin M, Hamlin JL, Bekiranov S. Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early- and late-firing origins. *Genome Res.* 2013;23(11):1774–88.
  127. Mendenhall EM, Koche RP, Truong T, Zhou VW, Issac B, Chi AS, et al. GC-rich sequence elements recruit PRC2 in mammalian ES cells. *PLoS Genet.* 2010;6(12):1–10.
  128. Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, Mikkelsen TS, et al. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.* 2008;4(10).
  129. Delgado S, Gómez M, Bird A, Antequera F. Initiation of DNA replication at CpG islands in mammalian chromosomes. *EMBO J.* 1998;17(8):2426–35.
  130. Prioleau MN. CpG Islands: Starting blocks for replication and transcription. *PLoS Genet.* 2009;5(4):4–5.
  131. Sequeira-Mendes J, Díaz-Uriarte R, Apedaile A, Huntley D, Brockdorff N, Gómez M. Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genet.* 2009;5(4).
  132. Smith OK, Kim R, Fu H, Martin MM, Lin CM, Utani K, et al. Distinct epigenetic features of differentiation-regulated replication origins. *Epigenetics and Chromatin.* 2016;9(1):1–17.
  133. Lubelsky Y, Prinz JA, DeNapoli L, Li Y, Belsky JA, MacAlpine DM. DNA replication and transcription programs respond to the same chromatin cues. *Genome Res.* 2014;24(7):1102–14.
  134. Picard F, Cadoret JC, Audit B, Arneodo A, Alberti A, Battail C, et al. The

- Spatiotemporal Program of DNA Replication Is Associated with Specific Combinations of Chromatin Marks in Human Cells. *PLoS Genet.* 2014;10(5).
135. Marks AB, Fu H, Aladjem MI. Regulation of replication origins. *Adv Exp Med Biol.* 2017;1042:43–59.
  136. Long H, Zhang L, Lv M, Wen Z, Zhang W, Chen X, et al. H2A.Z facilitates licensing and activation of early replication origins. *Nature [Internet].* 2020;577(7791):576–81. Available from: <http://dx.doi.org/10.1038/s41586-019-1877-9>
  137. Clément C, Orsi GA, Gatto A, Boyarchuk E, Forest A, Hajj B, et al. High-resolution visualization of H3 variants during replication reveals their controlled recycling. *Nat Commun [Internet].* 2018;9(1). Available from: <http://dx.doi.org/10.1038/s41467-018-05697-1>
  138. Wang Y, Long H, Yu J, Dong L, Wassef M, Zhuo B, et al. Histone variants H2A.Z and H3.3 coordinately regulate PRC2-dependent H3K27me3 deposition and gene expression regulation in mES cells. *BMC Biol.* 2018;16(1):1–18.
  139. Voigt P, Tee WW, Reinberg D. A double take on bivalent promoters. *Genes Dev.* 2013;27(12):1318–38.
  140. Bae S, Lesch BJ. H3K4me1 Distribution Predicts Transcription State and Poising at Promoters. *Front Cell Dev Biol.* 2020;8(May):1–11.

```
1 Appendix – 3:
2
3 import subprocess
4 import matplotlib.pyplot as plt
5 import numpy as np
6 import collections
7 import csv
8 from tqdm.autonotebook import tqdm, trange
9
10 HL_15 = "DIN124A106"
11 LL_15 = "JON91A17"
12 HL_3H = "JON91A13"
13 LL_3H = "JON91A9"
14
15 threshold = 0.00003
16 w=500
17
18 chromosomes = [[ 1, 248956422], [2, 242193529], [3, 198295559], [4,
19 190214555], [5, 181538259], [6, 170805979],
20 [7,159345973],
21 [8,145138636],
22 [9,138394717],
23 [10,133797422],
24 [11,135086622],
25 [12,133275309],
26 [13,114364328],
27 [14,107043718],
28 [15,101991189],
29 [16,90338345],
30 [17,83257441],
31 [18,80373285],
32 [19,58617616],
```

```

33 [20,64444167],
34 [21,46709983],
35 [22,50818468],
36 ['X',156040895],
37 ['Y',57227415]]
38
39 def get_bam_readcounts(file, chr, start, stop):
40     out = subprocess.Popen(['./bam-readcount', '-f',
41 '/home/souradip/igv/genomes/seq/hg38.fa.gz',
42
43 './bam/rep1/'+file+'.uniquely.mapped.pcr.dup.rm.bam',
44
45     'chr' + str(chr) + ':' + str(start) + '-' +
46     str(stop)
47     ],
48     stdout=subprocess.PIPE,
49     stderr=subprocess.STDOUT)
50
51     stdout, stderr = out.communicate()
52     result = stdout.decode("utf-8").split("\n")[1:]
53     return result
54
55 def get_bam_totalreads(file):
56     result = 0
57     with
58     open('./bam/rep1/'+file+'.uniquely.mapped.pcr.dup.rm.bam.count') as
59     f:
60         result = int(f.readlines()[0].strip())
61     return result
62
63 def parse_origin(line):
64     line = line.strip().split("\t")
65     return { 'chr': line[0], 'start': line[1], 'stop': line[2] }
66
67 def get_hl_ll_ratio_offset_width(hl_file, ll_file, origin, offset, w):

```

```

67     hl_15      =      get_bam_readcounts(hl_file,      origin['chr'],
68 int(origin['start']+offset, int(origin['start']+offset+w)

69     ll_15      =      get_bam_readcounts(ll_file,      origin['chr'],
70 int(origin['start']+offset, int(origin['start']+offset+w)

71     hl = 0
72     ll = 0
73
74     for f in hl_15:
75         if len(f.split("\t")) > 2:
76             hl += int(f.split("\t")[3])
77     hl /= get_bam_totalreads(hl_file)
78
79     for f in ll_15:
80         if len(f.split("\t")) > 2:
81             ll += int(f.split("\t")[3])
82     ll /= get_bam_totalreads(ll_file)
83
84     return ( hl ) - ( ll )
85
86
87
88 def smooth(y, box_pts):
89     box = np.ones(box_pts)/box_pts
90     y_smooth = np.convolve(y, box, mode='same')
91     return y_smooth
92
93 buffer = collections.deque(maxlen=40)
94 smoothed = np.array([])
95
96 detected = []
97 # chromosomes[chr][1]
98

```

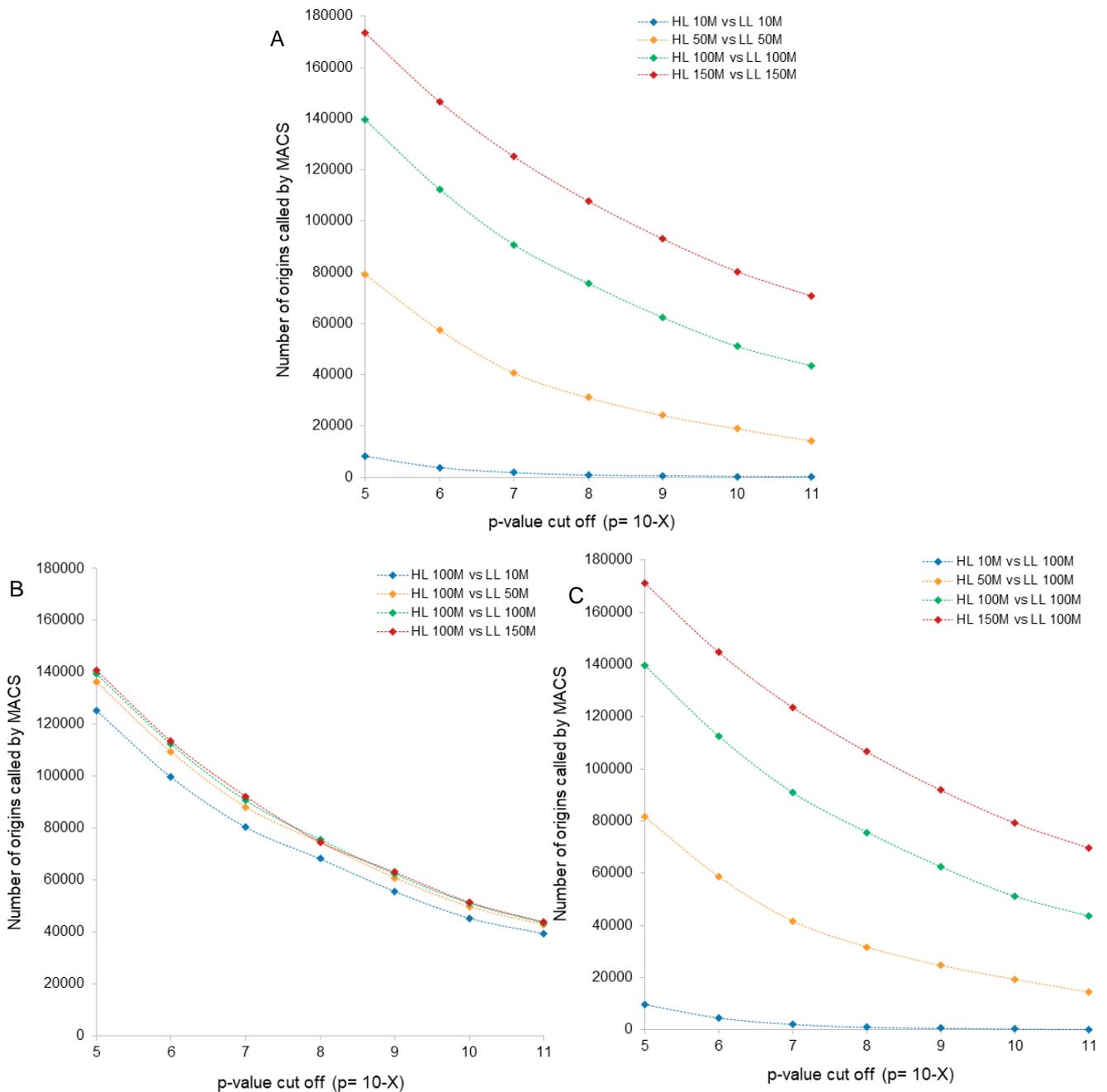
```

99 chr = 0
100 for chr in trange(len(chromosomes), desc="chromosomes"):
101     buffer = collections.deque(maxlen=40)
102     smoothed = np.array([])
103
104     detected = []
105     start_pos = 0
106     end_pos = chromosomes[chr][1]
107
108     with
109 open('dec_2020/rep1_chr'+str(chromosomes[chr][0])+'_15mins_thres_3_win
110 dow_500.csv', 'a+', newline='', buffering=1) as csvfile:
111         wtr = csv.writer(csvfile, delimiter=' ', quotechar='|',
112 quoting=csv.QUOTE_MINIMAL)
113         signal = False
114         averaging = []
115         for position in tqdm(range(start_pos, end_pos, w),
116 desc="progress"):
117             line = { "chr": str(chromosomes[chr][0]), "start":
118 str(position) }
119
120             value = get_hl_ll_ratio_offset_width(HL_15, LL_15, line, 0,
121 w)
122
123             buffer.append(value)
124             smoothed = smooth(buffer, 19)
125             if signal:
126                 averaging.append(value)
127
128             if buffer[-1] >= threshold:
129                 if not signal:
130                     signal = True
131                     detected.append([chromosomes[chr][0], position,
132 position, -1])
133             else:

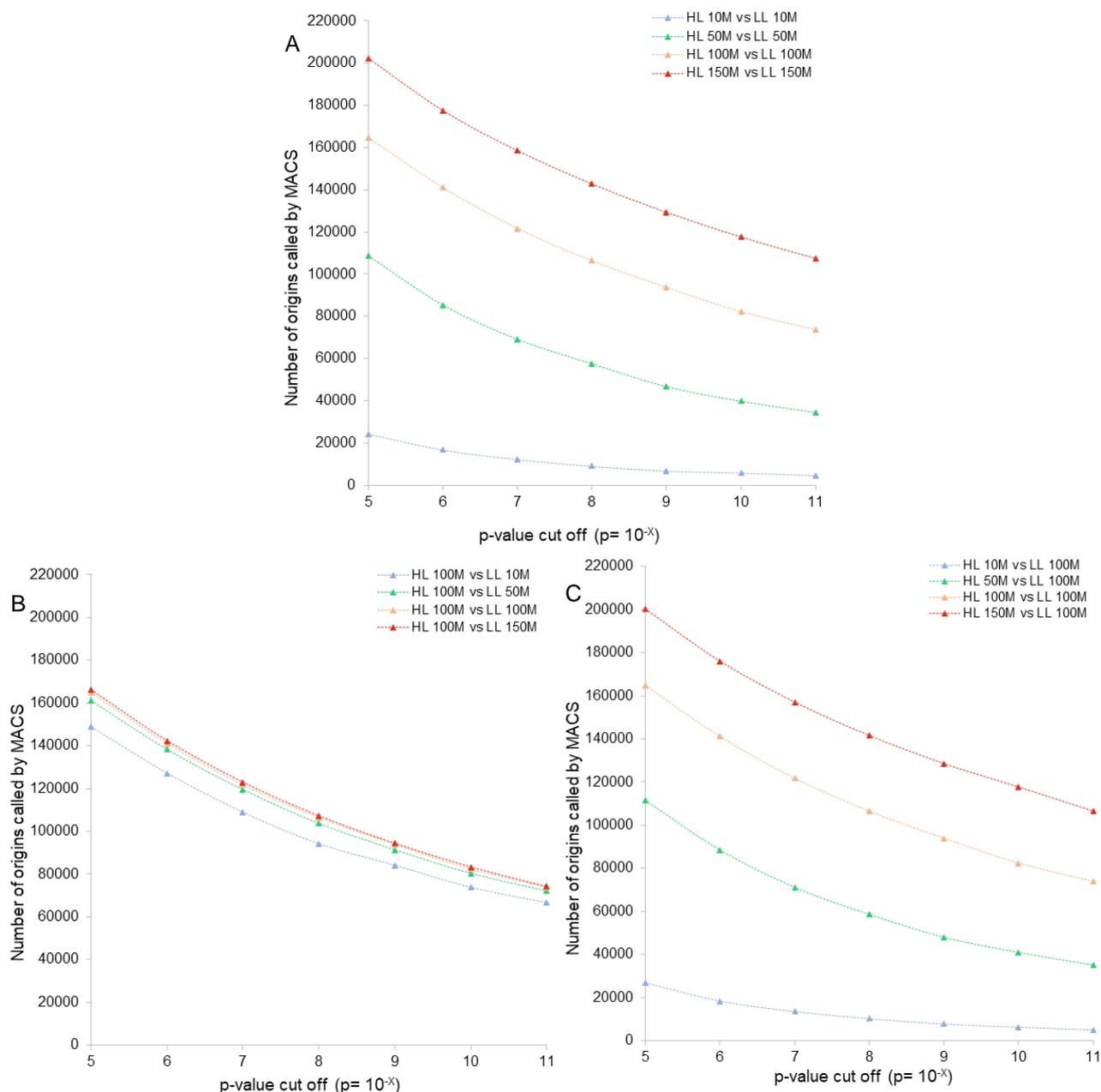
```

```
134         if signal:
135             signal = False
136             detected[-1][2] = position
137             detected[-1][3] = np.mean(np.array(averaging))
138             averaging = []
139         for x in detected : wtr.writerow (x)
140
141     #print(detected)
```

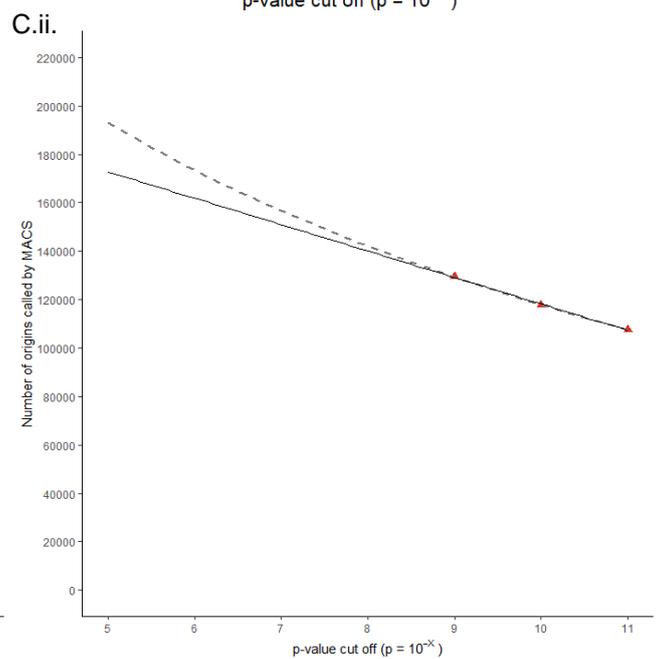
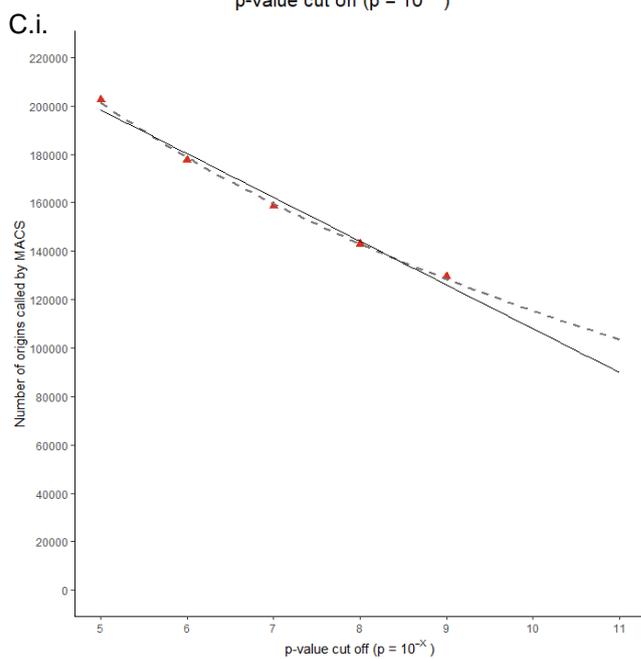
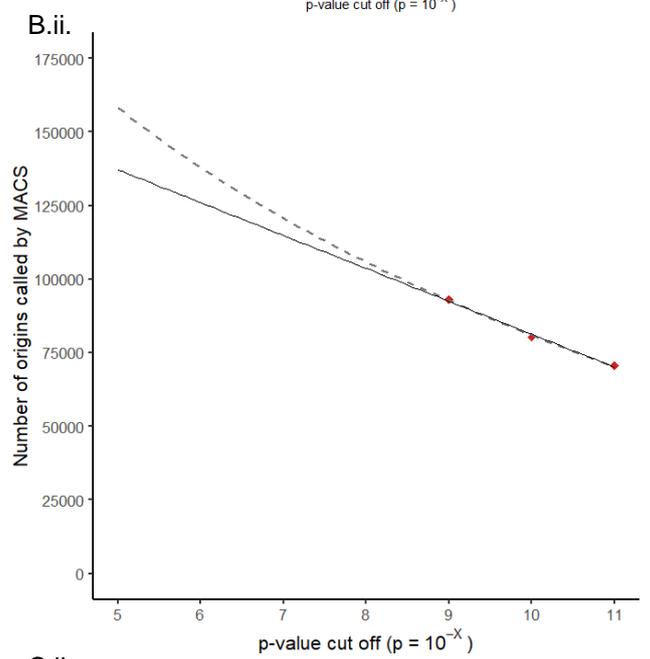
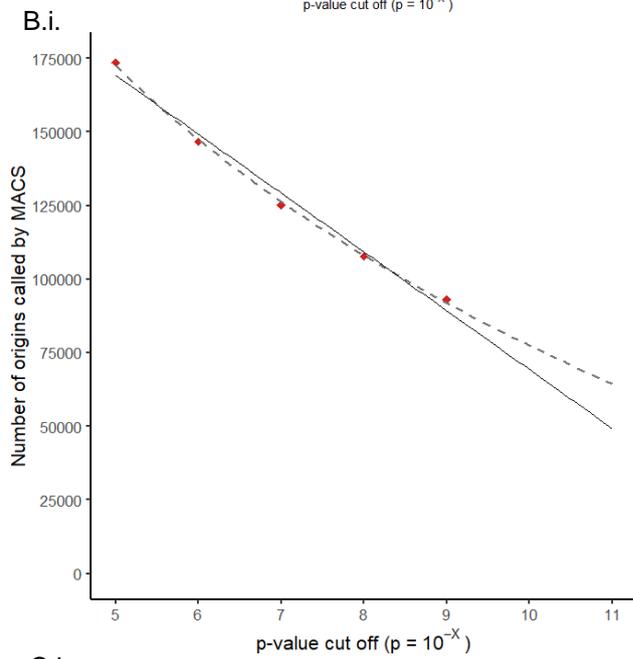
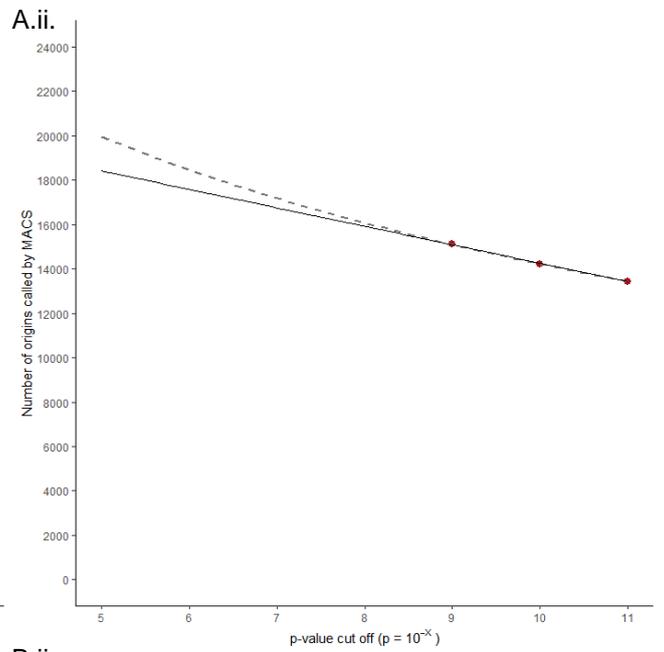
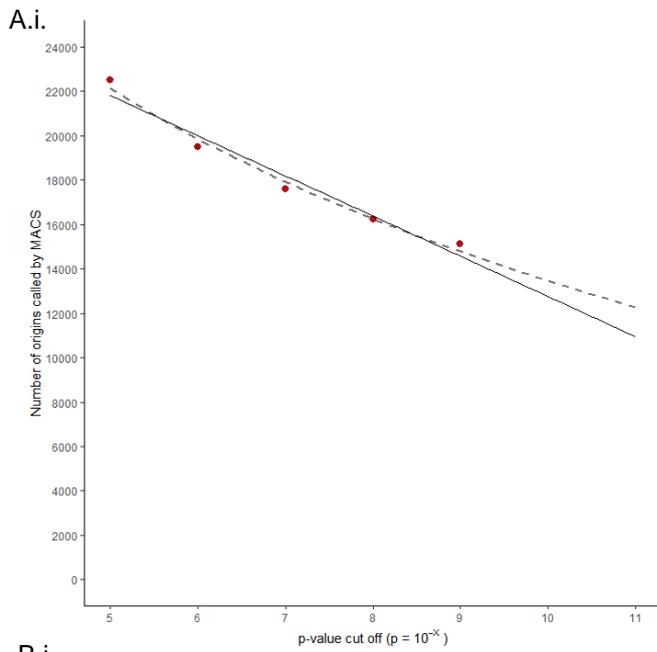
## Appendix – A4:



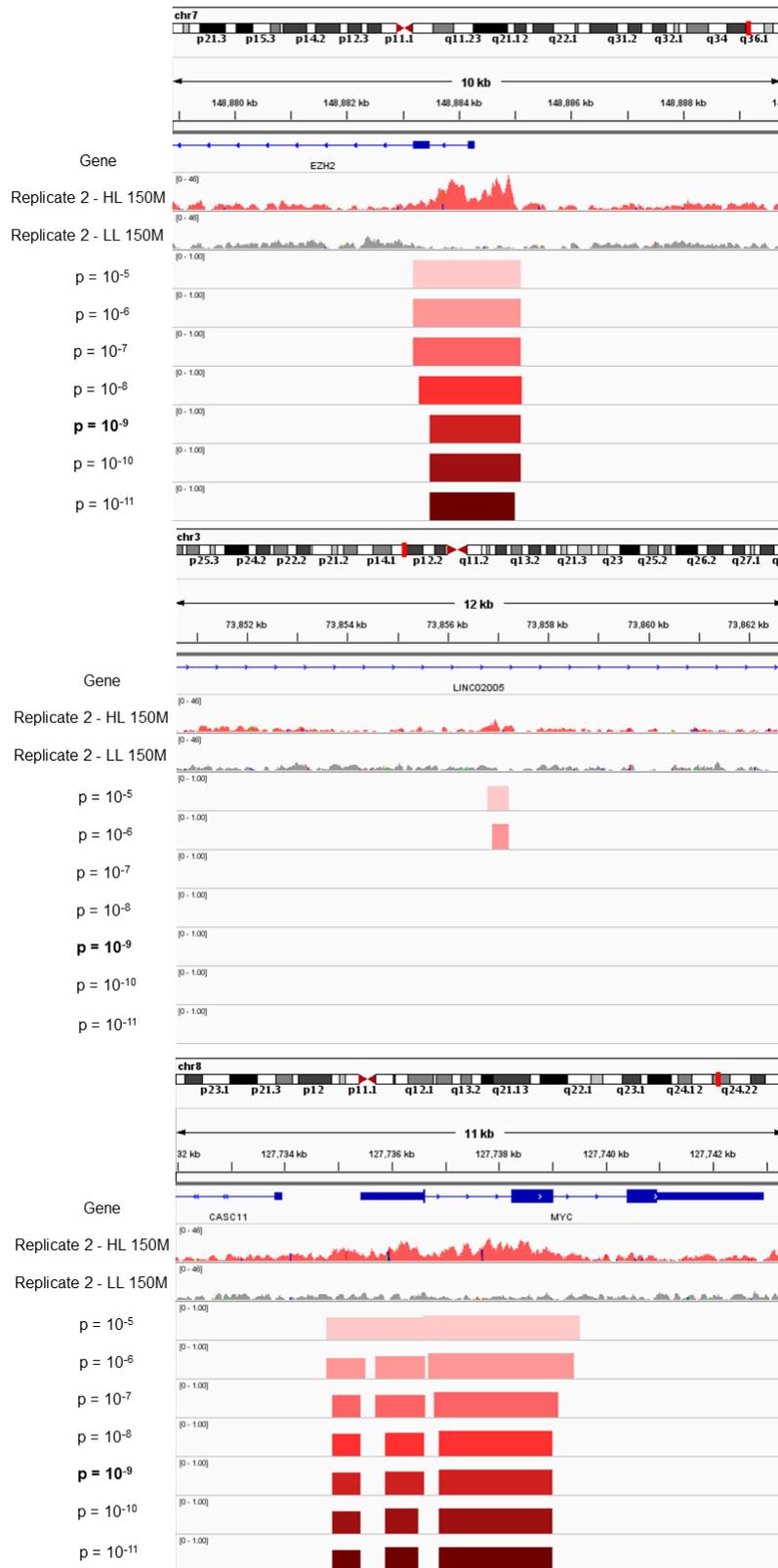
**Figure A4.1:** The sequenced files generated for the replicated HL and unreplicated LL of replicate 2 were of differing read numbers. Files with defined numbers of reads were generated for both the HL and LL and several titrations were performed. Origins were called using MACS peak caller at increasingly stringent p cut-off values (determined at which point the HL was significantly enriched above LL), using HL and LL data with differing total read numbers. (A) showed the “matched-pairs” titrations where the number of reads in HL = that of LL. The following conditions were shown: HL 10M vs LL 10M (blue), HL 50M vs LL 50M (green), HL 100M vs LL 100M (orange) and HL 150M vs 150M (red). (B) showed the “fixed HL” titrations where the number of reads in HL = that of LL. The following conditions were shown: HL 100M vs LL 10M (blue), HL 100M vs LL 50M (green), HL 100M vs LL 100M (orange) and HL 100M vs 150M (red). (C) showed the “fixed LL” titrations where the number of reads in HL = that of LL. The following conditions were shown: HL 10M vs LL 100M (blue), HL 50M vs LL 100M (green), HL 100M vs LL 100M (orange) and HL 150M vs 100M (red).



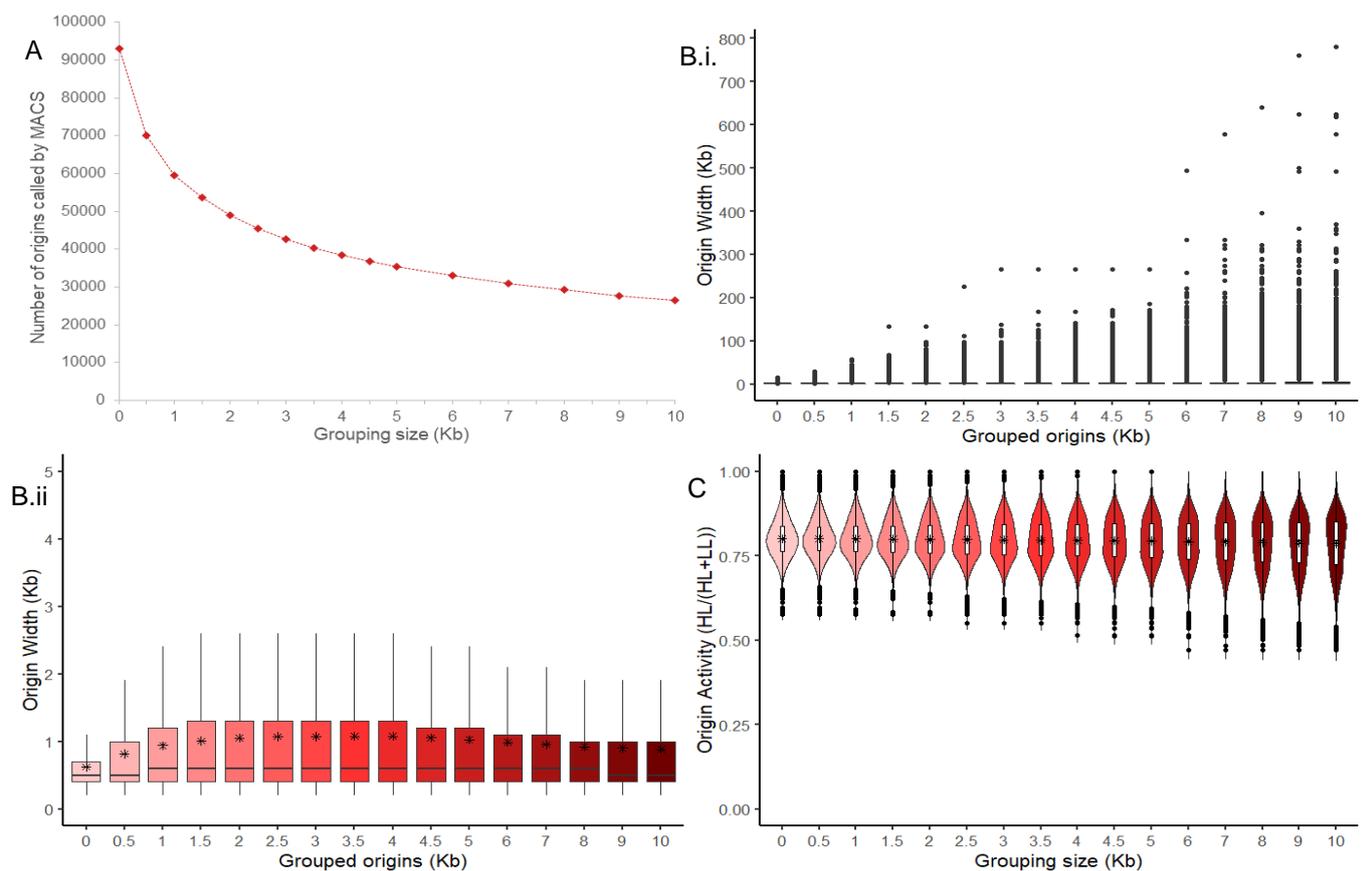
**Figure A4.2:** The sequenced files generated for the replicated HL and unreplicated LL of replicate 3 were of differing read numbers. Files with defined numbers of reads were generated for both the HL and LL and several titrations were performed. Origins were called using MACS peak caller at increasingly stringent p cut-off values (determined at which point the HL was significantly enriched above LL), using HL and LL data with differing total read numbers. (A) showed the “matched-pairs” titrations where the number of reads in HL = that of LL. The following conditions were shown: HL 10M vs LL 10M (blue), HL 50M vs LL 50M (green), HL 100M vs LL 100M (orange) and HL 150M vs 150M (red). (B) showed the “fixed HL” titrations where the number of reads in HL = that of LL. The following conditions were shown: HL 100M vs LL 10M (blue), HL 100M vs LL 50M (green), HL 100M vs LL 100M (orange) and HL 100M vs 150M (red). (C) showed the “fixed LL” titrations where the number of reads in HL = that of LL. The following conditions were shown: HL 10M vs LL 100M (blue), HL 50M vs LL 100M (green), HL 100M vs LL 100M (orange) and HL 150M vs 100M (red).



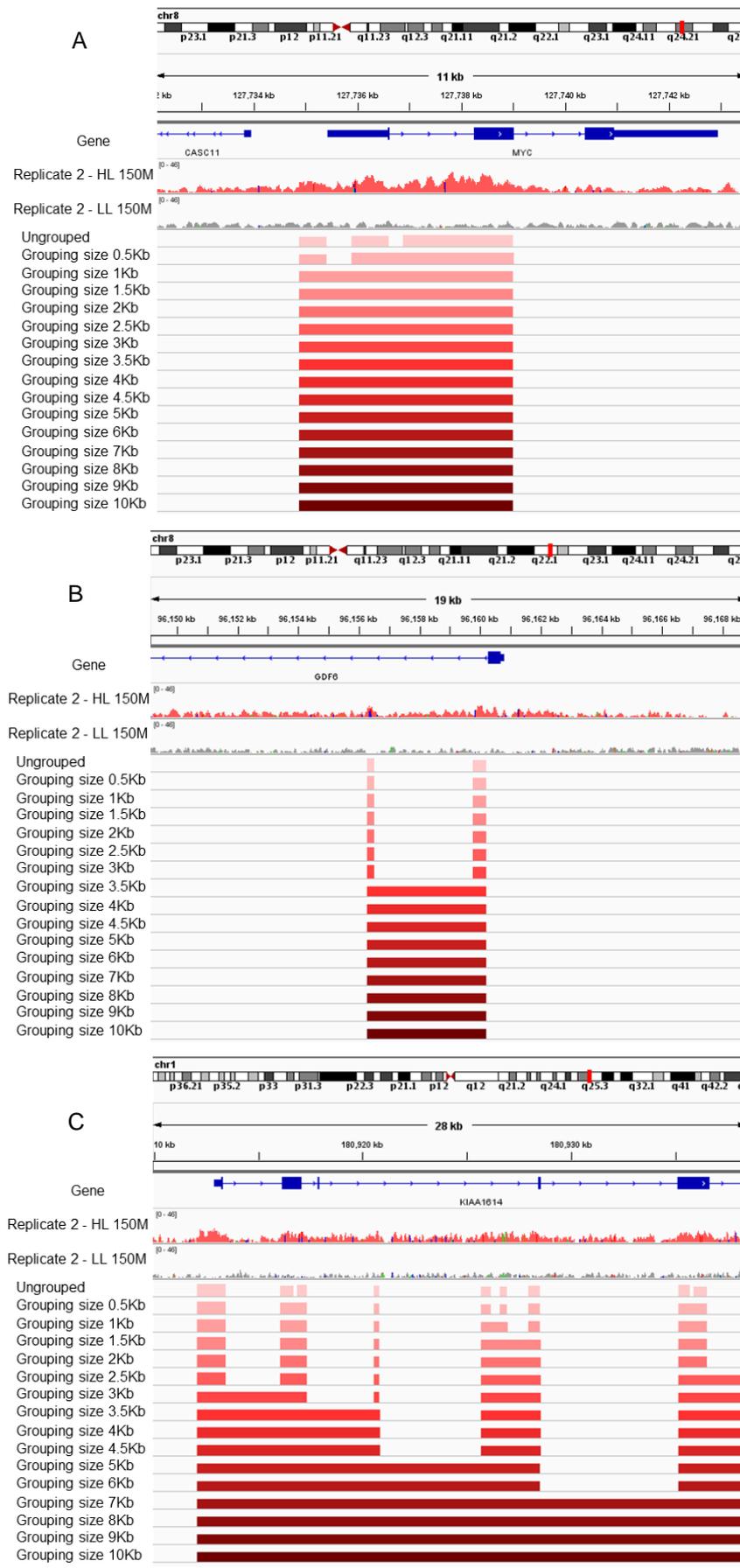
**Figure A4.3:** The numbers of origins called by MACS peak caller at increasingly stringent p cut-off values, where  $10^{-5}$  was the most lenient/least stringent and  $10^{-11}$  was the most stringent, for replicates 1 (A - dark red circle), 2 (B - cherry red diamond) and 3 (C - dark orange-red triangle); for all of the replicates, the read numbers for HL and LL were 150 million reads (replicate 1 LL was ~140 million reads as this was the size of the original file)). For each replicate, the least stringent portion of the data sets ( $10^{-5}$  to  $10^{-9}$ ) (i) fitted to both an exponential trendline (in the  $10^{-9}$  to the  $10^{-5}$  direction; a logarithmic trendline in the  $10^{-5}$  to  $10^{-11}$  direction) (grey dashed), and a linear trendline (black solid). The more stringent portion of the data sets ( $10^{-9}$  to  $10^{-11}$ ) (ii) were fitted to both an exponential trendline (in the  $10^{-9}$  to the  $10^{-5}$  direction; a logarithmic trendline in the  $10^{-5}$  to  $10^{-11}$  direction) (grey dashed), and a linear trendline (black solid).



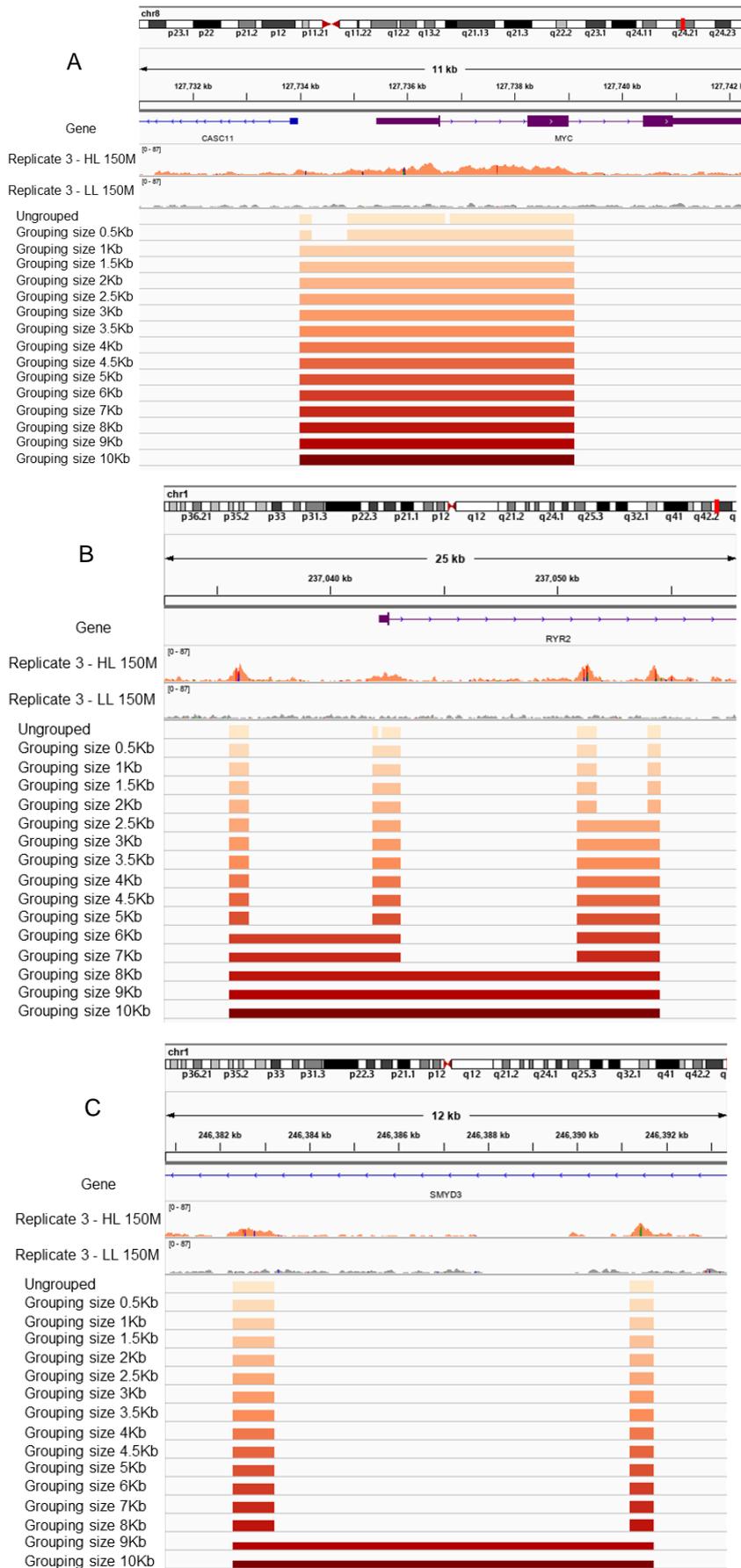
**Figure A4.4:** The IGV images show the mapped sequencing data generated from the replicate 2 replicated HL DNA (cherry red), the corresponding replicate 2 unreplicated LL DNA (grey), the reference genes (purple) and the corresponding origins called by MACS peak caller at increasingly stringent p cut-off values, where  $10^{-5}$  was the most lenient/least stringent and  $10^{-11}$  was the most stringent, for replicates 2 (gradient of cherry red colours) (read numbers for HL and LL were 150 million reads). The chromosome and position (bright red marker on the chromosome) on the chromosome were also indicated above the HL and LL profiles. The examples of called origins at different p cut-off values shown here are those found at the EZH2 promoter (A), LINC02005 gene body (B) and the MYC promoter (C). The selected p cut-off value of  $10^{-9}$  was highlighted in bold.



**Figure A4.5:** The numbers of origins called by MACS peak caller at a p cut-off value of  $10^{-9}$  (read numbers for HL and LL were 150 million reads), where origins were either ungrouped or grouped where origins were within 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 6, 7, 8, 9 and 10Kb of one another (aka grouping size). (A) showed the number of origins at found with these grouping parameters for replicate 2. (B) showed the widths of the replicate 2 origins, at the grouping parameters, for both the overall widths for all origins (i) and the origins with widths of up to 5Kb (ii) to highlight the interquartile ranges, medians and means (highlighted with \*). (C) showed the origin activities of the replicate 2 origins, at the grouping parameters, where the interquartile ranges, outliers and medians were indicated with an overlaying boxplot. The means were indicated with an \*.

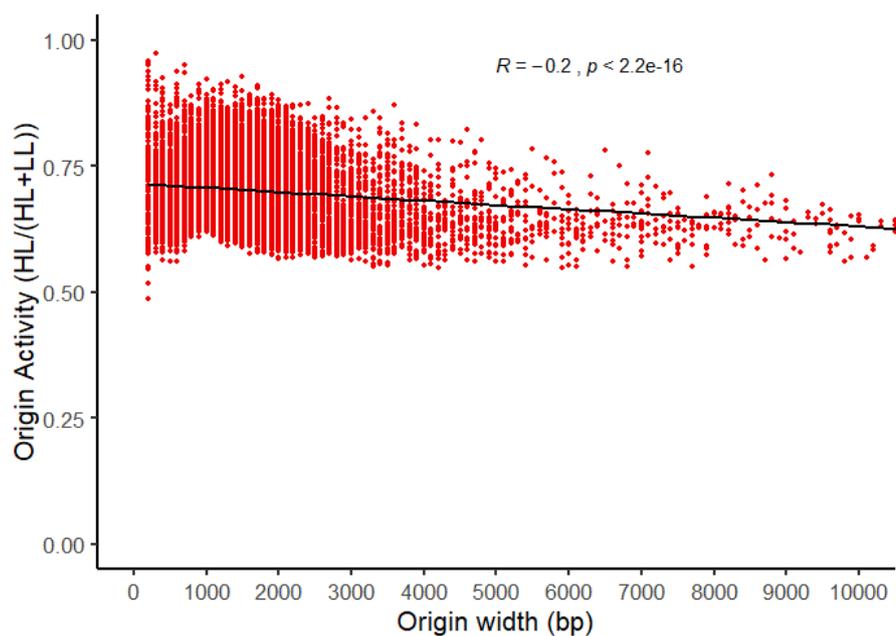


**Figure A4.6:** The IGV images show the mapped sequencing data generated from the replicate 1 replicated HL DNA (cherry red), the replicate 2 unreplicated LL DNA (grey for all replicates) (HL 150 million reads, LL 150 million reads) and the reference genes (purple). The chromosome and position (bright red marker on the chromosome) on the chromosome were also indicated above the HL and LL profiles. The origins (MACS peak caller,  $p = 10^{-9}$ ) identified from the grouping titration was also shown, from ungrouped origins to grouping sizes of 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 6, 7, 8, 9 and 10Kb (origins with those distances were grouped together and classed as 1 larger origin). (A) showed the example of the replication region/zone found at the MYC promoter, which demonstrated that grouping was required. (B) showed the example of origin(s) found at a gene promoter and the corresponding gene body, which highlighted that large grouping sizes ( $>3.5$ Kb) were inappropriate. (C) showed the example of origins found at gene body region, which highlighted that grouping size greater than 1Kb was inappropriate.

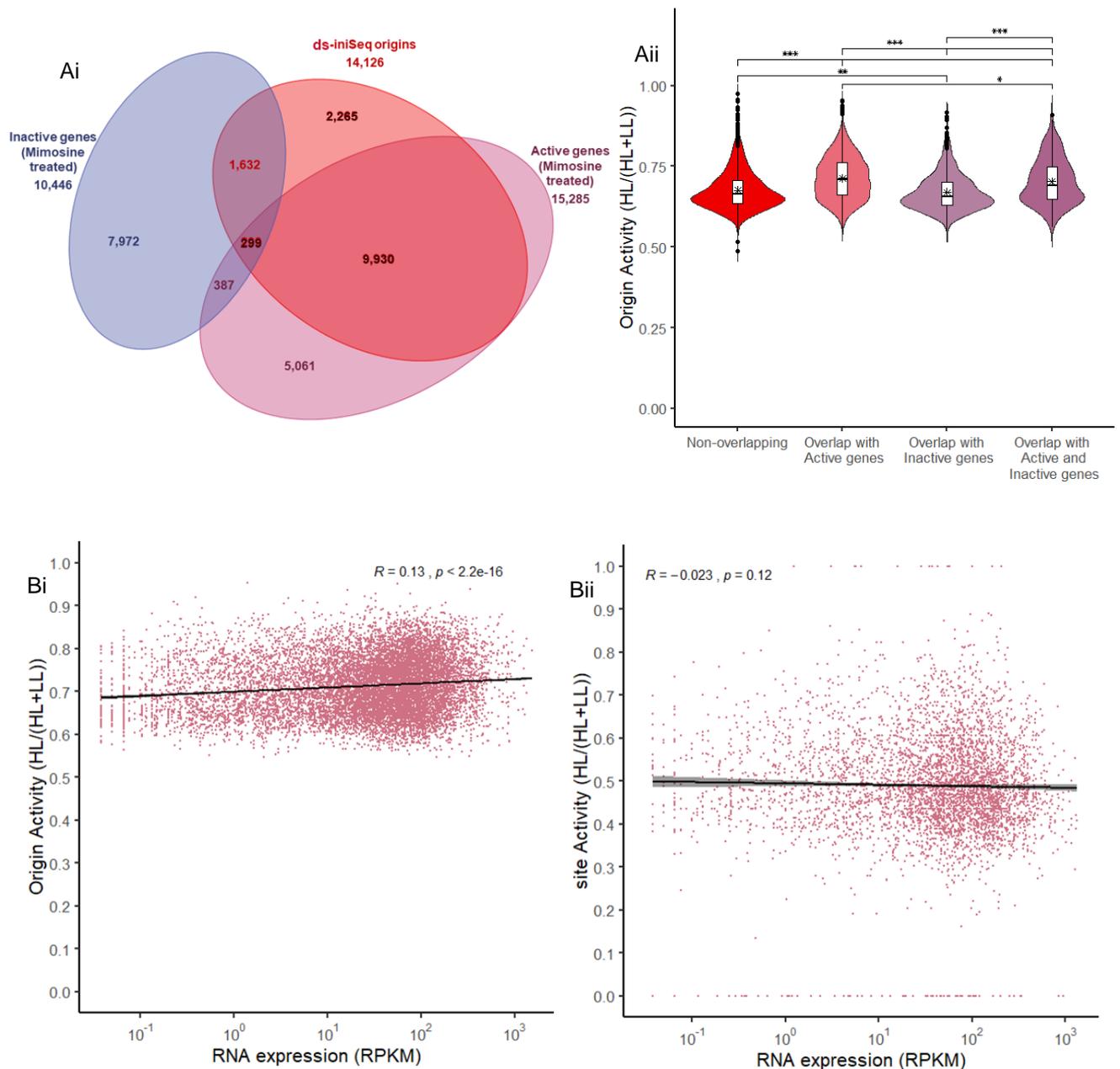


**Figure A4.7:** The IGV images show the mapped sequencing data generated from the replicate 3 replicated HL DNA (orange-red) the replicate 3 unreplicated LL DNA (grey for all replicates) (HL 150 million reads, LL 150 million reads) and the reference genes (purple). The chromosome and position (bright red marker on the chromosome) on the chromosome were also indicated above the HL and LL profiles. The origins (MACS peak caller,  $p = 10^{-9}$ ) identified from the grouping titration was also shown, from ungrouped origins to grouping sizes of 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 6, 7, 8, 9 and 10Kb (origins with those distances were grouped together and classed as 1 larger origin). (A) showed the example of the replication region/zone found at the MYC promoter, which demonstrated that grouping was required. (B) showed the example of origin(s) found at gene body and its neighbouring intergenic area, which highlighted that large grouping sizes (>2Kb) were inappropriate. (C) showed the example of origin(s) found at gene body region, which highlighted that the largest grouping sizes (9Kb & 10Kb) was inappropriate.

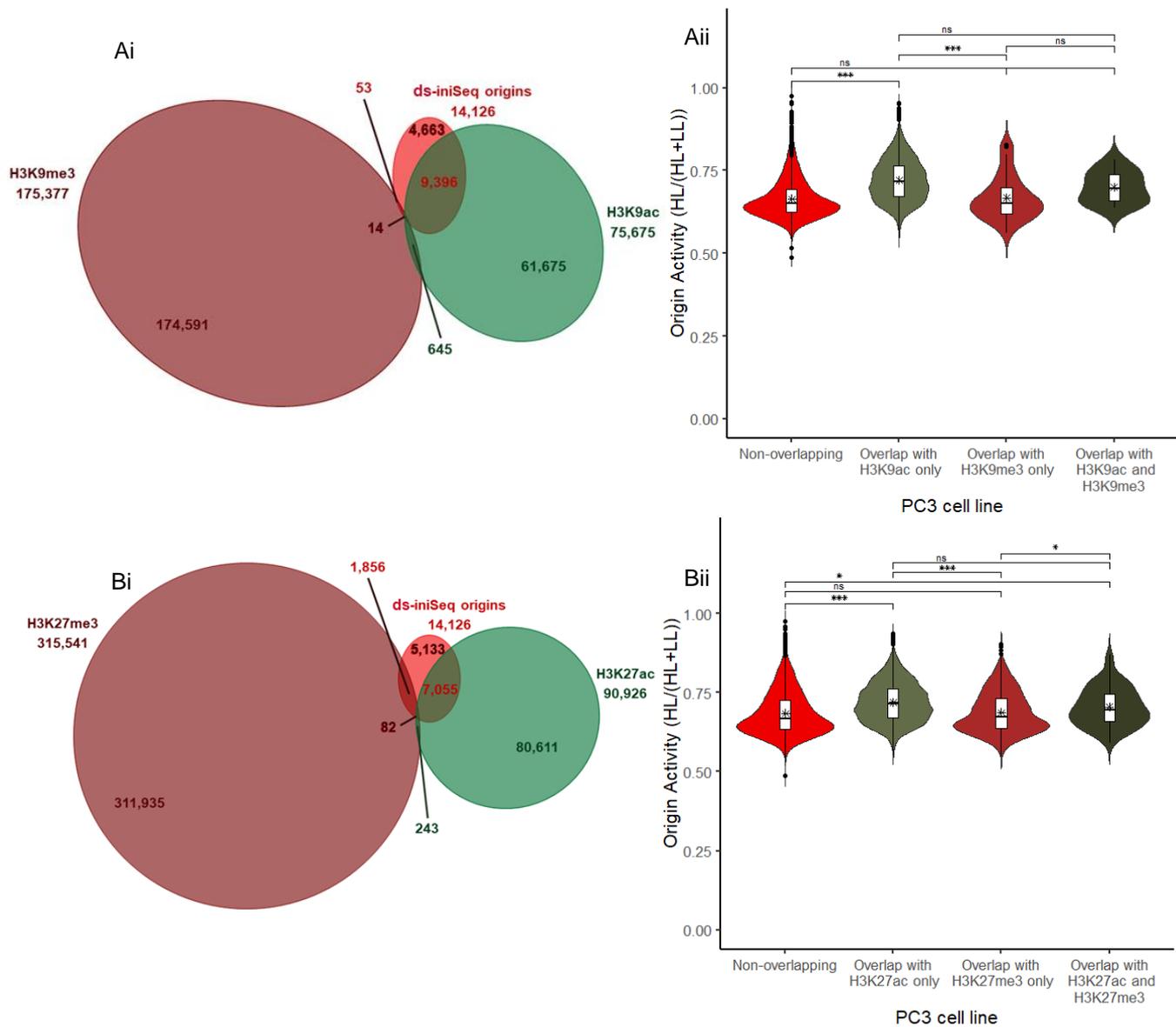
## Appendix – A5:



**Figure A5.1:** The scatter of the origin widths (up to 10Kb) and the corresponding origin activities. A linear regression, Pearson test for correlation and ANOVA test for significance were conducted on the whole data set (full range of widths) and shown on this plot.



**Figure A5.2:** The called origins found in replicate 1, that overlap with the called origins in replicates 2 and 3 were compared to the combined (and unique) annotated genes (blue purple) where transcript isoforms were merged (ensemble). (A) Mimosine-treated RNA expression was determined from the quantification (SeqMonk RNA-seq quantitation pipeline) of total RNA-seq of mimosine-treated EJ30 cells. Active (Reads per kilobase of transcript per million reads (RPKM) < 0) and inactive (RPKM = 0) genes were determined from the mimosine RNA expression. (i) The 3 way overlap of active genes (mauve) and inactive (blue purple) with the ds-iniSeq origins (red). (ii) The origin activity of the ds-iniSeq origins that; did not overlap with active or inactive genes (“non-overlapping; overlapped with only active genes; overlapped with only inactive genes; and that overlapped with both active and inactive genes. The means are indicated with an \*. An ANOVA and subsequent Tukey’s post-hoc test were performed to assess significance; the Tukey’s test results are shown on the plot and \*\*\* indicates  $p < 0.001$ , \*\* indicates  $p < 0.01$  and \* indicates  $p < 0.05$ . (Bi) Mimosine RNA expression was determined from the quantification of total RNA-seq of EJ30 cells treated with mimosine (to synchronise cells in late G1) for 24 hours. The origin activity of ds-iniSeq origins overlapping genes was plotted against the mimosine RNA expression of the corresponding gene. (Bii) The origin activity of ds-iniSeq random sites overlapping genes was plotted against the mimosine RNA expression of the corresponding gene. (B) A linear regression, Pearson test for correlation and ANOVA test for significance were conducted and are indicated on each plot.



**Figure A5.3:** The H3K9 and H3K27 histone positions are both acetylated and trimethylated which are associated with early and late firing replication origins respectively. (Ai) The 3 way overlap of the early replication associated H3K9ac (PC3 cell line; dark green) and the late replication associated H3K9me3 (PC3 cell line; dark red) with ds-iniSeq origins (red). (Aii) The origin activity of the ds-iniSeq origins that; did not overlap with H3K9ac or H3K9me3 (“non-overlapping”); overlapped with only H3K9ac; overlapped with only H3K9me3; and that overlapped with both H3K9ac and H3K9me3. (Bi) The 3 way overlap of the early replication associated H3K27ac (PC3 cell line; dark green) and the late replication associated H3K27me3 (PC3 cell line; dark red) with ds-iniSeq origins (red). (Bii) The origin activity of the ds-iniSeq origins that: did not overlap with H3K9ac or H3K27me3 (“non-overlapping”); overlapped with only H3K9ac; overlapped with only H3K27me3; and that overlapped with both H3K27ac and H3K27me3. (All ii) The means are indicated with an \*. An ANOVA and subsequent Tukey’s post-hoc test were performed to assess significance; the Tukey’s test results are shown on the plot and \*\*\* indicates  $p < 0.001$ , \* indicates  $p < 0.05$  and ns indicates “not significant”.