UNIVERSITY OF CAMBRIDGE

ROBINSON COLLEGE

Role of DNA replication timing in gene expression and chromatin organisation

MIGUEL DINIS MONTEIRO DOS SANTOS

This thesis is submitted for the degree of Doctor of Philosophy

January 2022

Declaration of authorship

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text.

All bioinformatic analyses presented here are my own work.

All experiments presented here are my own work except for:

- Replication profiles and RNA-Seq libraries from Chapter 3 which were generated by Dr. Mark Johnson, a post-doc from the lab. My project started with the bioinformatic analysis of these datasets.
- Origin deletion experiments from Chapter 3 which were performed by Dr. Mark Johnson. I was responsible for the analysis of the replication profiles.

It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text.

It does not exceed the prescribed word limit for the Degree Committee for the Faculty of Biology.

Abstract

Role of DNA replication timing in gene expression and chromatin organisation

Miguel Dinis Monteiro dos Santos

Eukaryotic DNA is duplicated according to an evolutionary conserved temporal pattern. This pattern of DNA replication is altered during development and differentiation and can be dysregulated in cancers. While temporal changes in genome duplication are associated with altered transcription and chromatin organisation, it is still unknown whether DNA replication timing (RT) is a cause or a consequence of cellular fate changes. During my thesis I used a conditional system to perturb DNA RT in a single cell cycle in budding yeast, in combination with whole-genome sequencing techniques such as replication profiles, RNA-Seq and MNase-Seq in order to understand the biological importance of a defined pattern of genome replication and the impact on the genome structure and function. Overall, dramatic changes in gene expression, chromatin structure and transcription-factor (TF) binding events were observed, and a significant number of genes affected are involved in differentiation processes such as sporulation. While some differentially expressed genes showed significant chromatin changes, there were also examples where this was not the case, as well as genes with changes in chromatin and no changes in expression, which illustrates the complex nature of the relationship between RT, gene expression and chromatin. Differential TF binding events explained some of the observed changes, supporting a role for RT to maintain the correct TF binding dynamics during S-phase. Additionally, the fact that budding yeast origins are defined by specific sequences allowed the local modulation of RT which showed a direct effect of RT on gene expression. Altogether, the work generated during this thesis provides insight into the complex relationship between replication timing, gene expression and the chromatin landscape.

For my family,

Acknowledgments

This work wouldn't be possible without the help and support from several people, to whom I am forever grateful. I want to thank my supervisor, Dr. Philip Zegerman, for giving me the opportunity to be part of this amazing story. Your mentorship and support were invaluable. A special thanks to Mark, who taught me every single experiment present in this thesis, and with whom I worked side by side since my first day in the lab. This project is as much mine as it is yours. Thank you for your guidance and friendship. I want to thank every single member of the Zegerman lab (past and present) for creating a fun and stimulating environment, namely: Esther, Kang Wei, Fiona, Geylani, Manuela, Vincent and Florence. I want to thank the core staff of the Gurdon Institute, with a special note to Sylviane, for keeping us on track, Al and the IT team, Kay for taking care of the sequencing of our samples, Charles for help and advice with analysis and everyone in the administrative team for making our lives so much easier. A special thanks to the whole Media Kitchen, for the phenomenal work, including my dear friend Miguel for taking care of my crazy out of hours requests and for the fun we had together. To Lukas, who ended up teaching me more than what I taught him, and contributed with an immense amount of work which was determinant for the direction this project took. I also want to thank BBSRC and the Cambridge Trust for funding my PhD – without this funding I wouldn't be able to study in Cambridge and my college, Robinson, for being the first home I had in Cambridge and a place where I made friends for life: Kyriakos, Stavros, Osama and Chiara. You were with me at the start and made the most difficult year I had in Cambridge better than what I could ever hope for. To Shannon and Theresa, for the fun, laughs, parties and constant support. To Lisa, for always having time for me. To Andreas, for being one of the most loyal friends I have. To Pantelis and Andria, for never saying no. We went through a lot together, and I know I can always count on you. To Torcato, Pipas, Migas and Dudu, for welcoming me in their family, which is now also part of mine. I will never think of Cambridge without thinking of you. To Mar for invaluable help with printing and binding and for being part of a new start.

I wouldn’t be where I am if it wasn’t for my oldest and closest friends, which I consider my family. There are no words to describe how much I like you and depend on you. Maria, you are one of the few friends with whom I have spent more than half of my life and you keep surprising me every day. To Rita, my oldest friend. To Inês J., Sofi and Margarida. To Inês M. and Debanjan. Marco, when I grow up I want to be like you (shh). João Simão, I don’t think it is physically possible not to be happy around you. Filipe, I know you would follow me until the end of the world. Patrick, és como um irmão. Meu querido Xico, the stories we have together feel like 5 lifetimes. Bia, you are my sunshine and my other half. Malhadas, there is nothing left to say between us. I hope you know the amount of respect and love I have for you. My dear friends, I live through you and for you.

To my Family: I have to be the luckiest person ever for growing up and being surrounded by you. To my dear cousins, aunts and uncles, who make me always want to come back home: Tita, Rui, Mia, Zé, Mixa, Joaquim, Kikó, Vi, Carla, Rafinha, Mó, Sofia, Bela, Manel, Inês, Ana, Kiki, Kika, Pi, Só, Nuno, Joana, Paulo, Francisco, Afonso, Lina, João, Diogo, Sara. And to my new family: Maria Helena, Zé António, Tiago, Rafa, António, Laura, Sofia, Jorge, Matilde, Simão, Daniel, James, Kristin. To my godfather Oca, who treats me like a son and always believed in me. To my dear grandparents, Avô João, Avó São and Avô Arlando. To my biologist grandmother Avó Susete: I wish you were here to see this! To my Luke, for the unconditional love. To my Sister Maria for being a constant source of inspiration and pride, I will take care of you all my life. To my Brother Tiago, my role model and the person who put me where I am now. You were by my side since the day I was born and I miss you every day we are apart. To Andreia and Gaia for taking care of you. To my Mom and Dad for making me the person I am, I owe everything to you. You showed (and keep showing) me what it means to love and feel loved. Finally, to my dear Diana, the love of my life. You brought a whole new meaning to my life, which is now more yours than mine. Everything I do is for you and every day I love you more and more. We will be together until the end, I promise.

I could be born 20 times and I would always choose every single one of you.

I am happy and complete because of you.

Table of Contents

List of tables and figures

List of abbreviations and symbols

DNA	DeoxyRibonucleic Acid
S-phase	Synthesis phase
RT	Replication Timing
S. cerevisiae	*Saccharomyces cerevisiae*
ARS	Autonomously Replicating Sequences
ACS	ARS Consensus Sequence
2D gel	Two-dimensional gel
bp	Base-pairs
AT	Adenine/Thymine
ORC	Origin Recognition Complex
GC	Guanine/Cytosine
TAD	Topologically Associated Domain
G1	Gap 1 phase
NGS	Next-generation sequencing
T_{rep}	Time of Replication
ssDNA	Single-stranded DNA
HU	Hydroxyurea
dNTP	Deoxyribose nucleotide triphosphate
Rad53	RADiation sensitive 53
Δ	Delta (gene deletion)
BrdU	Bromodeoxyuridine
ChIP	Chromatin Immunoprecipitation
MCM	MiniChromosome Maintenance
GINS	*Go-Ichi-Nii-San* (Japanese for 5-1-2-3)
α-factor	Alpha-factor (budding yeast mating pheromone)
Pre-RC	Pre-replicative complex
ATP	Adenosine Triphosphate
NDR	Nucleosome depleted region
Cdc6	Cell Division Cycle 6
Cdt1	Cdc10 dependent transcription
S-CDK	Cyclin-dependent kinase
DDK	Dbf4-dependent kinase
Cdc45	Cell Division Cycle 45
Sld3	Synthetic lethal with dpb11-1
Dpb11	DNA Polymerase B subunit 11
Sld2	Synthetic lethal with dpb11-1
CMG	Cdc45/Mcm2-7/GINS
SSDDCS	Overexpression Sld2-Sld3-Dbf4-Dpb11-Cdc45-Sld7
Sld7	Synthetic lethal with dpb11-1

Rpd3	Reduced potassium dependency
Sir2	Silent information regulator
Rif1	RAP1-interacting factor
TF	Transcription-factor
Fkh	Forkhead homolog
TSA	Trichostatin A
RNA	Ribonucleic acid
tRNA	Transfer RNA
mRNA	Messenger RNA
H3	Histone 3
Rtt109	Regulator of Ty1 transposition
DNase	Deoxyribonuclease
Hi-C	High throughput chromatin conformation capture
CTCF	CCCTC-binding factor
ERCE	Early replicating control elements
kb	Kilobase
dNTPs	Deoxynucleoside triphosphate
Sml1	Suppressor of Mec1 lethality
rDNA	Ribosomal DNA
rtQTL	Replication timing quantitative trait loci
C. elegans	*Caenorhabditis elegans*
URA3	URAcil requiring
JAK2	Janus kinase 2
RNA-Seq	RNA sequencing
MNase-Seq	Micrococcal nuclease digestion sequencing
ml	milliliter
µl	microliter
LRT	Likelihood ratio test
PCA	Principal Component Analysis
DANPOS	Dynamic Analysis of Nucleosome Position and Occupancy by Sequencing
TSS	Transcription start site
ROC	Receiving operator characteristic
IGV	Integrative Genomics Viewer
FACS	Fluorescence-activated cell sorting
DE	Differentially expressed
FDR	False discovery rate
GO	Gene Ontology
ACF	Autocorrelation function
EMG	Early meiotic genes
YetFaSCo	Yeast Transcription Factor Specificity Compendium
FIMO	Find Individual Motif Occurrences

Chapter 1 – Introduction

1.1 - What is the replication timing programme?

DNA replication is the process used by all living cells to make an identical copy of the genome prior to cell division. This fundamental process ensures that daughter cells inherit the complete set of genetic instructions needed for their growth, development and function. Due to its importance, eukaryotic DNA replication consists of several highly regulated steps, which involves the action of a complex protein network. These ensure that DNA replication is coordinated with the cell cycle, monitored by checkpoints and coupled to chromatin inheritance[1].

Although the whole genome has to be replicated perfectly before cell division, not all parts of the genome replicate at the same time during S-phase. Eukaryotic DNA replication starts at discrete regions of the genome called origins of replication and some origins initiate replication (i.e. “fire”) before others. This temporal pattern of origin firing is known as the replication timing (RT) programme and the first observations supporting such a programme go back to 1960s, when Taylor observed that Chinese hamster cells incorporated [^{3}H]-thymidine in different chromosomal regions at different times during S-phase[2] and Lima de Faria observed that euchromatin is replicated before heterochromatin[3]. This defined pattern of origin firing is conserved from unicellular eukaryotes such as budding yeast *Saccharomyces cerevisiae*[4] to metazoans[5]. Its evolutionary conservation suggests a biological significance for organisms, but its exact importance has remained elusive[6].

Many studies have shown that RT is correlated with genomic features such as gene expression and chromatin structure[6], and while these support a mechanistic link, it has been hard to determine cause and effect (is RT the main determinant of gene expression and chromatin patterns, or just a consequence of them?). Moreover, RT is regulated during cellular differentiation, allowing the identification of different cells types solely based on their replication profiles[7,8]. Finally, RT is disrupted in cancer and its dysregulation leads to genomic instability and increased mutagenesis[9].

1.2 - Unravelling the replication timing programme: from single origins to the whole-genome

How can one study origin firing? First of all, one needs to know where origins are located across the genome. Similar to the principles governing transcription, it was postulated that regions of replication initiation would be defined by specific sequences, where the machinery responsible for replication initiation could bind and start replicating DNA. Such genomic locations were identified in budding yeast using a plasmid maintenance assay. In this assay, different pieces of genomic DNA are introduced into plasmids which are unable to replicate autonomously. Sequences of DNA that supported plasmid replication were named Autonomously Replicating Sequences (ARS)[10]. A few years later, using a two-dimensional (2D) agarose gel electrophoresis assay, it was shown that ARSs act as replicators in their native chromosomal locations[11]. Since then, these sequences have been extensively characterised: all ARSs in budding yeast are characterised by a 100 to 200 bp sequence consisting of a 11-bp AT-rich domain called ARS consensus sequence (ACS)[12], together with further sequence elements that act as a binding site for the origin recognition complex[13] (ORC, described in more detail in subsequent sections). Surprisingly, only *S. cerevisiae*, some other yeast species[14] and defined regions of metazoan genomes such as the *Drosophila* chorion gene locus[15] have such clear sequence-defined origins.

In the distantly related fission yeast *Schizosaccharomyces pombe* origins have little sequence conservation, but consist of discrete 1kb AT-rich regions (about 70%) which distinguishes them from the rest of the genome[16]. In human cell systems, the plasmid maintenance assay has shown that any tested DNA fragment, if long enough, could provide a plasmid with replicative ability, suggesting that human origins have low sequence specificity[17]. Moreover, many groups have tried to identify such consensus sequences in metazoans as well as genetic and epigenetic features defining origins of replication, with limited success. Many genomic features have been shown to be associated with the location of origins of replication in metazoans, such as GC content, histone marks, topologically associated domains (TADs) boundaries and protein-DNA binding patterns, among others[18], but none of

these has been shown to be absolutely required nor sufficient to define DNA replication start sites.

The fact that origins are defined by exact sequences in *S. cerevisiae* allows unprecedented tractability of origin firing patterns. The first experiments monitoring origin firing kinetics used a variation of the Meselson-Stahl dense isotope transfer method[19]. McCarrol and Fangman determined the approximate time of replication of centromeres and telomeres in budding yeast by growing cells in cultures with dense medium until their DNA was fully labelled with the heavy isotope (heavy-heavy). Then, cells were arrested in G1 and released in a synchronous S-phase in medium with a light isotope. Samples were collected in different time-points and un-replicated DNA (heavy-heavy) was separated from replicated (heavy-light) by centrifugation[20]. This experiment showed for the first time that, in budding yeast, centromeres are early replicating while telomeres are late replicating. While human telomeres were found to be mostly late replicating[21] (but it was also shown that they can be replicated throughout S-phase[22]), the early replication of centromeres seems to be a yeast specific phenomenon[22,23]. Since then, the combination of 2D gels with dense isotope transfer allowed the characterisation of firing kinetics of several individual origins and some complete chromosomes in yeast, providing small scale / low resolution profiles of the replication dynamics of an eukaryotic genome[24–26].

The advance of genomics allowed the DNA replication field to move from the analysis of small groups of origins to a whole-genome view of replication kinetics[27]. Using the dense isotope transfer method together with DNA microarrays, Raghuraman et al. generated the first genome-wide replication profile of the yeast genome[4]. By comparing the relative abundance of different genomic regions in the replicated (heavy-light) and un-replicated (heavy-heavy) DNA fraction during S-phase, the authors were able to generate replications profiles for each chromosome where peaks represent regions that replicate before their neighbouring sequences (origins) and valleys termination zones. The taller a peak is, the earlier the origin fires[4] (Fig. 1.1).

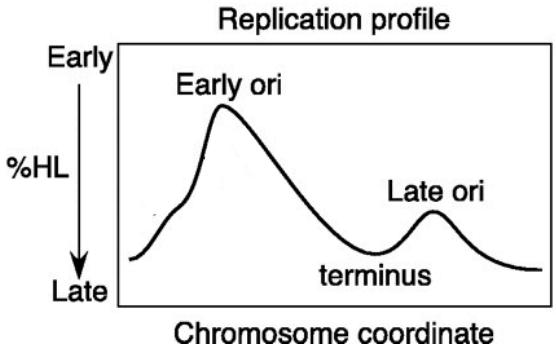


Figure 1.1 – Generation of replication profiles using the dense isotope method and microarrays. The relative abundance of specific genomic sequences present in the un-replicated versus replicated DNA at different S-phase time-points (represented as %HL) is plotted as a function of the chromosome coordinate to generate replication profiles. Peaks correspond to the location of origins and the taller a peak is, the earlier the origin fires. Valleys represent termination zones. Figure modified from Raghuraman et al.[4]

The development of next-generation sequencing (NGS) largely replaced microarrays, allowing an unbiased (i.e., not relying on previously known sequences) high-throughput genome-wide view of replication dynamics[28]. Using this approach, DNA collected in different S-phase time-points is deep sequenced and the copy number ratio of S-phase samples versus a non-replicated control (usually a G1 sample) is calculated[29]. From these profiles, one can estimate the time of half-maximal replication for each genomic region, which is known as time of replication or T_{rep} (Fig. 1.2).

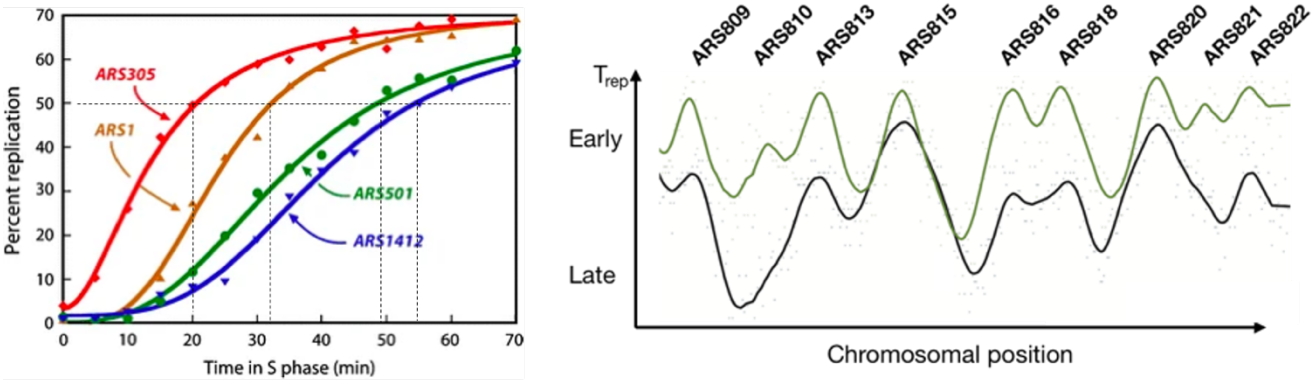


Figure 1.2 – From copy number ratios to time of replication (T_{rep}) profiles. Left – The kinetics of replication of 4 yeast origins, as a ratio of replicated to un-replicated DNA in each S-phase time-point (figure modified from Raghuraman and Brewer 2009[27]). Dashed lines indicate how T_{rep} can be determined from the replication curves. **Right** – T_{rep} profile for yeast chromosome VIII. T_{rep} values are plotted along the chromosome coordinates and smoothed. The locations of known ARS were overlayed over the plot and these align with peaks in the profile, as expected. This data is from replication profiles generated in this thesis and will be discussed in subsequent chapters, used to illustrate a representation of T_{rep} profiles. The green curve represents a strain where replication timing was advanced genome-wide and the black curve the corresponding control strain. ARS – Autonomously Replicating Sequences.

Additional whole-genome techniques have increased our knowledge of DNA replication dynamics, which I will describe briefly. The mapping of single-stranded DNA (ssDNA) in cells starting S-phase in the presence of hydroxyurea (HU) has also allowed the identification of origins that fire early in S-phase. HU delays S-phase by decreasing the dNTP pool[30], so early origins are the only ones that can fire and accumulate peaks of ssDNA[31]. Rad53 is the kinase responsible for blocking late origin firing in HU[32], so mapping of ssDNA in *rad53*Δ mutants allows the identification of early and late origins[31]. An alternative method for the identification of sites of replication initiation is the incorporation of nucleotide analogues such as bromodeoxyuridine (BrdU) followed by immunoprecipitation and sequencing or hybridization[33,34]. A complimentary approach to the sequencing of replicated DNA involves the chromatin immunoprecipitation (ChIP) of components of the replication machinery that bind origins, such as ORC and MCM[35] or GINS[36]. These footprints allow the identification of origin locations across the genome, and are particularly useful for the identification of dormant origins which usually do not fire but are bound by replication proteins.

These methods have been applied to different organisms under different experimental conditions[5,7,8], but most pioneering studies were performed in budding yeast, due to its simple genetic manipulation and wide-range of methods for synchronisation in different cell cycle stages, such as conditional mutants, elutriation centrifugation and the mating pheromone α-factor[37]. Despite the immense amount of information provided by these genome-wide approaches, the fact that a population of cells is being analysed has to be considered when interpreting the results. The sigmoidal shape of origin firing kinetics (Fig. 1.2 Left), rather than a sharp step function suggests that the same origin does not fire at the same time in all cells in a population. Perhaps not surprisingly, single-cell and single-molecule approaches demonstrated the heterogeneous nature of origin firing across cell populations, which I will discuss in the next section.

1.3 - Dissecting the replication timing programme: efficiency vs timing and the stochastic nature of origin firing

Replication profiles generated from whole-genome methods represent the behaviour of most cells in the population (Fig. 1.1 and 1.2), and do not necessarily account for heterogeneity within the population. This is one of the major disadvantages of population based methods and the reason why important aspects of the RT programme are still not completely understood.

Analysis of the replication kinetics of yeast chromosome VI by DNA combing, a technique that allows the visualisation of replication patterns in single DNA molecules by BrdU incorporation followed by fluorescence microscopy, has shown that no two molecules had the exact same pattern[38], suggesting that each cell has its own unique replication profile. Moreover, the authors compared their single molecule results with published population based replication profiles, by averaging the profiles of 105 single chromosome VI molecules. Briefly, the single molecules were divided in bins of equal size and each bin was attributed a numerical value depending on whether this bin was replicated or not. The average of the 105 values for each bin can be considered as a proxy of the probability that each region is replicated in the entire population. When the authors overlayed this probability map with published T_{rep} profiles, they found a surprisingly good overlap between the averaged profile and population based profiles[38]. This single molecule study helped with the biological interpretation of the results from population based replication profiles and suggests that these profiles represent probability maps of origin firing.

The fact that population based profiles represent the probability of origins firing still does not allow the differentiation between two different conceptual aspects of origin kinetics: origin efficiency (i.e. the proportion of cells in which it fires)[6] and origin firing time. The two are partially correlated, and while firing can be considered an innate characteristic of origins, efficiency on the other hand is dependent on the effect of passive replication initiated by neighbouring origins. For example, two different scenarios could explain an origin with late T_{rep} in a cell population:

1 – The origin fires very efficiently in late S-phase across most cells or;

2 – The origin fires early but is inefficient and as such is passively replicated by incoming forks in most cells.

Therefore, T_{rep} values from whole-genome replication profiles represent a combination of cells where the origin has actively fired and cells where the origin is passively replicated. This is particularly relevant in the larger genomes of metazoans, where origins are not sequence defined and peaks in replication profiles represent the heterogeneous firing of several initiation sites (whose position varies slightly between individual cells) with similar timing that form mega-base sized replication domains[6].

Altogether, single molecule and population studies suggest that despite the defined order of replication genome-wide, origin firing is stochastic at the single cell level[28,39–41]. This is further supported by mathematical models based on whole genome profiles[42,43], which also suggest that origin firing is independent from the firing of nearby origins. The advance of single-cell sequencing techniques[44,45], as well as more advanced single molecule approaches[39,46], provided the ultimate confirmation that origin firing is stochastic at the single-cell level, despite the stability of the temporal order of origin firing. If origin firing is indeed a stochastic process, how can some origins replicate consistently early or late during S-phase in most cells, a feature that is conserved across organisms[6]? This evolutionary conservation suggests that the temporal order of origin firing has a fundamental biological role, but there is no obvious reason for such a temporal programme simply to accomplish duplication of the genome. During the next sections I will describe how origin firing (and consequently the RT programme) is regulated, as well as briefly discuss how stochastic firing of individual origins can be reconciled with defined temporal patterns of replication of different genomic regions, providing some insight on how evolution has shaped replication dynamics.

1.4 – Cell cycle regulation of DNA replication

The first step necessary for DNA replication is the formation of the pre-replicative complex (pre-RC) during late mitosis / early G1 at potential origins, which is called licensing[1]. A licensed origin is biochemically competent to initiate DNA replication, i.e. all components required for replication initiation are loaded onto chromatin. Interestingly, eukaryotes license more origins than the ones needed to replicate the genome, which represents a strategy to guarantee that the entire genome is duplicated when replication forks are impeded[47] (this process will be described in more detail in the next section). Upon helicase activation during the start of S-phase, origins fire and start replicating the genome. The licensing of new origins is inhibited during S-phase to avoid re-replication.

1.4.1 – Licensing: preparation of origins for replication initiation

Licensing starts with the binding of the origin recognition complex (ORC) to origins, which was initially identified as a multiprotein complex that binds origins in yeast in an ATP-dependent manner[13]. There are more ORC binding consensus sequences across the genome than actual ORC binding events, which is explained by the fact that origin's chromatin needs to be in an accessible state for efficient binding[48]. Origins are characterised by a nucleosome depleted region (NDR) which is flanked by well positioned nucleosomes, while ORC consensus regions which are not bound by ORC are depleted from nucleosomes but are not flanked by well positioned nucleosomes. These observations suggest that well defined chromatin architecture is important for origin definition and function[48]. The next licensing step involves the loading of 6 minichromosome maintenance proteins (Mcm2-7), which form the core of the replicative helicase[49]. Two additional proteins are required for helicase loading, Cdc6 and Cdt1. The current model in the field supports that ORC binds Cdc6 and this complex recruits Cdt1 and Mcm2-7[49]. After loading of the first helicase onto an origin, a second one is loaded with the opposite orientation, as origins are consistently bidirectional[50]. ORC, Cdt1, Cdc6 and Mcm2-7 form the pre-RC, which binds all potential origins in G1 (Fig. 1.3 top). However, only a subset of licensed origins will fire during S-phase, making helicase activation the limiting step for replication initiation.

1.4.2 – Initiation: activation of (a subset of) loaded helicases

As mentioned previously, not all licensed origins fire in any given cell cycle, making helicase activation, also termed replication initiation, the limiting step for the replication reaction. Helicase activation is regulated by a complex series of phosphorylation events, which are dependent on the action of two kinases: S-CDK (cyclin dependent kinase) and DDK (Dbf4-dependent kinase). DDK phosphorylates the loaded Mcm2-7[51], which drives the recruitment of Cdc45 and Sld3 to the Mcm2-7 double hexamer[52]. CDK then phosphorylates its two essential targets Sld2 and Sld3, allowing their interaction with Dpb11[53] which drives the recruitment of GINS to origins[54], completing the activation of the replicative helicase by formation of the CMG complex (Cdc45/Mcm2-7/GINS) (Fig. 1.3 bottom). DNA unwinding by the CMG is activated by Mcm10 and ATP hydrolysis[55], resulting on the movement of the two opposing CMG helicases on each DNA strand.

1.4.3 – How do organisms ensure replication once per cell cycle?

Licensing and helicase activation are temporally separated, in order to ensure that DNA replication occurs once and only once per cell cycle. In eukaryotes this separation is achieved by limiting helicase loading to late mitosis / early G1 phase, and helicase activation to S-phase[56]. By separating licensing from activation, cells ensure that origins that have fired are not relicensed until the next cell cycle. In yeast, this separation is entirely regulated by CDK, which blocks licensing through mechanisms involving phosphorylation of ORC, nuclear exclusion of Mcm2-7 and Cdc6 degradation[57] and as such restricts helicase loading to the G1 phase, when CDK concentration is low. On the other hand, helicase activation requires high S-CDK levels (as described above), and as such is restricted to S-phase (Fig. 1.3).

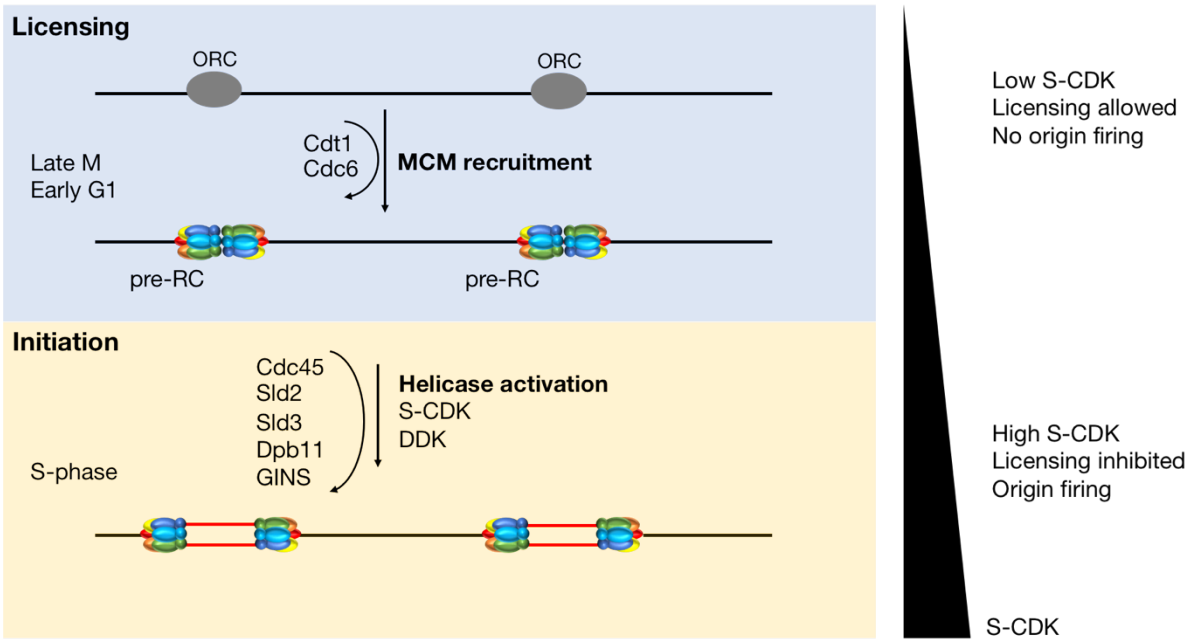


Figure 1.3 – DNA replication control during the cell cycle. **Top** – Licensing takes place during late mitosis / early G1 when CDK levels are low. All potential origins are "marked" with the origin recognition complex (ORC). Then, the MCM double hexamer is recruited by Cdt1 and Cdc6, completing the pre-replicative complex (pre-RC). **Bottom** – Initiation or helicase activation takes place during S-phase, when S-CDK levels increase. This step is regulated by a series of phosphorylation events which are dependent on S-CDK and DDK, allowing the recruitment of the factors that complete the active helicase (Cdc45, Sld2, Sld3, Dpb11 and GINS, among others). The high levels of S-CDK block further licensing during S-phase, avoiding re-replication and separating licensing and initiation to two non-overlapping cell cycle phases. See main text for detailed description.

1.5 – How do organisms ensure the timely replication of the entire genome?

As described previously, the entire eukaryotic genome has to be precisely replicated during S-phase, which is achieved through the action of bidirectional replication forks emanating from origins of replication distributed throughout the genome. During this process, replication initiation must not occur in a region that has been already replicated (re-replication) and no region should be left un-replicated.

I have described in the previous section how re-replication is avoided by separating licensing and initiation to non-overlapping cell cycle phases (Fig. 1.3) and in this section I will describe how organisms ensure a complete copy of the entire genome by leaving no un-replicated region behind.

1.5.1 – Dormant origins and origin distribution across the genome

Errors are inevitable during DNA replication, and if two converging forks irreversibly stall (double fork stall), the region between these two forks would be left un-replicated if there were no backup mechanisms present, because licensing of new origins is inhibited once S-phase has started (Fig. 1.3). Eukaryotic cells overcome this problem by licensing more origins during late mitosis / early G1 than the ones needed to replicate the genome[58], so these normally "dormant" origins can be used under replicative stress conditions. In mammalian cells for example, only approximately 10% of licensed origins are used during a normal S-phase[59]. These dormant origins are most likely just very inefficient origins which do not fire in most cell cycles and act as the backup to the complete replication of the genome when forks are stalled.

However, if no dormant origin is present between a double fork stall event, it would not be possible to replicate this region. Telomeres also represent problematic regions, because they are replicated by a single fork, so stalling of this fork (telomeric fork stall) would leave this region un-replicated if no dormant origins are present. Mathematical models supported by experimental data suggest that the probability of fork stalling events that would compromise the complete replication of the genome of yeasts is minimised by the distribution of origins of replication across the genome: origins are regularly spaced, large inter-origin gaps are minimised and the end-most origins are located closer to chromosome ends than expected by chance[60]. In the case of metazoans, which have significantly larger genomes compared to yeast, regularity of origin spacing is lost and larger gaps between adjacent origins are more common, so double fork stall events become almost inevitable genome-wide[61]. As such, organisms with larger genomes rely on post-replicative mechanisms to repair these errors[62].

However, the larger genomes of metazoans are organised in replication domains comprising several origins firing with similar timing, allowing for a better redistribution of resources under conditions that block or slow down fork progression ("divide and conquer mechanism")[59]. During replication stress conditions, checkpoint mechanisms redistribute resources quickly to finish replication within a domain (by firing dormant origins within the problematic domain) and inhibit initiation in domains that have not yet initiated until errors are resolved[59]. This organisation in replication domains and its implications for the RT programme will be discussed in subsequent sections.

1.5.2 - The random gap problem

As described previously, several studies support the stochastic firing of origins throughout S-phase. This could also lead to un-replicated gaps in the genome, as random origin firing throughout S-phase would occasionally lead to large un-replicated gaps which would take a long time to replicate – this is called the random gap problem (Fig. 1.4 top)[63]. One possible mechanism to avoid the generation of large gaps is the increase in origin efficiency as S-phase progresses, so origins in large un-replicated gaps are increasingly more likely to fire (Fig. 1.4 bottom)[63].

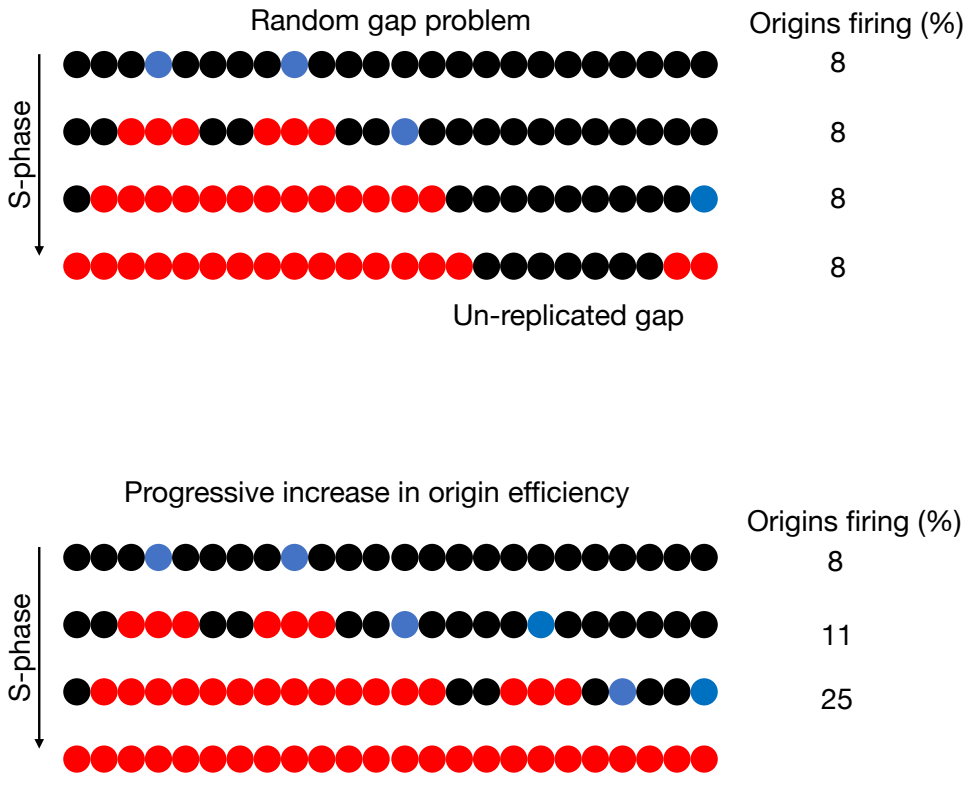


Figure 1.4 – The random gap problem. Top – Stochastic origin firing would occasionally lead to un-replicated gaps which would take a long time to replicate and delay S-phase completion. Black circles represent potential origins which can fire (blue circles) with 8% probability. Replication takes place bi-directionally, turning un-replicated origins (black) into replicated (red). Of the 24 potential origins, 2 fire during the first time-point (8% of total). However, in the second time-point there are only 18 un-fired origins, of which only one fires (~8% of total). By random chance, no origin on the right side has fired during early S-phase, leaving an un-replicated gap. **Bottom** – In this scenario, the efficiency of origin firing increases with S-phase progression. Most un-fired origins in late S-phase are at the un-replicated gap, making them more likely to fire as S-phase progresses and to finish replication in this region. Figure based on Rhind et al.[63]

A potential explanation for a progressive increase in origin efficiency during S-phase is the recycling of limiting initiation factors. If factors which are required for replication initiation are present in lower levels compared to the number of licensed origins, only the most accessible origins will fire at the start of S-phase. Once early origins have fired, these factors are released and fewer origins are left un-replicated, making these origins better suited to compete for the pool of initiation factors and more likely to fire in late S-phase. I will describe experimental data supporting this model in subsequent sections, as well as its implications to RT control.

Moreover, several studies support a model in which the timing of replication of an origin is dependent on the origin's ability to compete for diffusible initiation factors, and this ability is influenced by different aspects such as the chromatin environment and binding patterns of proteins not directly associated with the replisome machinery, which I will also discuss in the next section.

As such, it is possible to reconcile the stochastic firing of individual origins with a defined temporal pattern of replication through:

1 – The differential affinities of origins to limiting initiation factors and;

2 – The consequent increase in firing efficiency during S-phase progression

For example, the left side of figure 1.4 could illustrate an early replicating region formed by efficient origins, while the right side could illustrate a late replicating region formed by less efficient origins.

In sum, evolution has shaped the distribution and efficiency of origins so that the probability of errors is minimised and the entire genome is timely replicated. This process is regulated at the level of single origins in budding yeast[60] as described, while metazoans rely on a "divide and conquer" mechanism[59], by organising the genome in replication domains formed by several origins with similar timing.

1.6 - Regulation of licensing and initiation and implications for the replication timing programme

As described previously, the timing of origin firing represents a combination of the intrinsic probability of firing and the origin's chromosomal context. This suggests that RT is regulated at two different mechanistic levels: execution (i.e. the ability of origins to compete for replication activators, which is a proxy of their firing probability) and establishment (i.e., the factors that set the ability of origins to compete for initiation factors, and as such, set their firing probabilities). The two fundamental regulatory steps of DNA replication, helicase loading (licensing) and activation (initiation) (Fig. 1.3), are linked to the establishment and execution of the RT programme.

1.6.1 – Establishment of replication timing during licensing is dependent on origins' chromatin context

Both in budding yeast and metazoans, the signal for early or late replication is established in G1[64,65], when the pre-RC is formed. Several studies suggest that the establishment of RT is dependent on the regulation of the chromatin context of origins[6] and early origins tend to bind ORC for longer periods during the cell cycle[66] and have more MCM loaded[67], which increases their firing probability. Paradoxically, one study has shown that budding yeast origins with very high affinity to ORC *in vitro* tend to fire very late and in a small percentage of cells[68].

When ARS1 and ARS501, an early and late firing origin respectively, were swapped with each other, they both acquired the firing time of the replaced origin, rather than

keeping their own[69]. In contrast, a similar larger scale study has found that while some origins acquired the firing time of the chromosomal region they were inserted in, others kept their intrinsic firing time, suggesting that the chromatin context is not sufficient to define RT[70]. The fact that origins in budding yeast can be classified into two different classes depending on whether the surrounding chromatin environment affects ORC binding or not[71], supports this view.

A recent study has further compared origins with different ORC affinities and has found that origins with weak ability to recruit ORC are sensitive to MCM levels, and depletion of MCM leads to delays in their firing time[72]. The authors proposed a model where ORC activity becomes important when MCM levels are reduced, as origins with more active ORC will be able to load MCM more efficiently and consequently will not be affected by MCM depletion. Therefore, under physiological conditions, ORC levels have little effect on MCM loading. Most importantly, overexpression of MCM did not affect the RT programme, suggesting that MCM levels are not limiting for replication initiation[72].

A recent *in vitro* study has confirmed that nucleosome organisation at origins affect ORC binding and helicase loading[73] and one study in human cells suggests that ORC/MCM density is correlated with replication timing but is not the main determinant of replication initiation[74]. These studies indicate that ORC/MCM binding is regulated by the origin's chromatin landscape and has an impact on firing time for a subset of origins. Recently, it was shown that replication patterns in yeast, mouse and human broadly reflect MCM binding, but some exceptions exist[75].

As described in previous sections, more origins are licensed compared to the ones that fire, so helicase activation determines how many origins fire across the genome (and when they fire). Therefore, ORC and MCM affect both the establishment and the execution of the RT, through the interaction with rate limiting factors required for helicase activation.

1.6.2 – Execution of replication timing during initiation is dictated by an origin's ability to compete for limiting factors

Several studies have shown that the execution of the RT is primarily dictated by the ability of origins to compete for limiting initiation factors for helicase activation. There are two ways of experimentally modulating this competition: by changing the accessibility of origins to these factors or by directly perturbing their local activity[1].

Two seminal studies have demonstrated that the overexpression of a subset of low abundance initiation factors causes an advance in origin firing[76,77]. Mantiero et al. observed that the overexpression of CDK essential targets Sld2 and Sld3, together with overexpression of their binding partner Dpb11 and DDK regulatory subunit Dbf4 (SSDD strain) cause the early firing of late origins[76]. Moreover, the overexpression of SSDD in combination with overexpression of Cdc45 and the Sld3 interacting partner Sld7 (SSDDCS strain), causes a genome-wide advance in RT ([78] and Zegerman lab unpublished observations present in this thesis). Tanaka et al. obtained similar results by overexpressing Sld3, Sld7 and Cdc45, which caused the firing of late origins earlier in S-phase[77].

On the other hand, modulating the ability of origins to compete for these limiting factors has similar effects on RT. Yoshida et al. found that two histone deacetylases, Rpd3 and Sir2, affect RT in an opposing manner. While Rpd3 represses late origin firing, Sir2 is required for initiation of early origins, so *rpd3Δ* and *sir2Δ* have nearly identical replication profiles[79]. The authors proposed that these two proteins regulate the ability of origins to compete for initiation factors, as overexpression of limiting factors suppresses the initiation defects of *sir2Δ* mutants and *SIR2* deletion restores the replication programme of *rpd3Δ* cells[79]. This study suggests that altering the pool of origins which are able to compete for limiting factors has an effect on RT. Another study has shown that the increase in acetylation caused by Rpd3 deletion advances RT in yeast[33] and Mantiero et al. have also shown that SSDDCS overexpression in *rpd3Δ* cells allows the firing of dormant origins in addition to early firing of late origins[76]. Another example of differential regulated accessibility is the late replication of telomeres, which are early replicating when Sir3, a protein involved in heterochromatin formation, is mutated[80].

1.6.3 – *trans* factors affecting accessibility of origins to limiting factors

Additionally, the budding yeast centromeric and telomeric regions were used to demonstrate that *trans* factors mediating the local activity of limiting initiation factors affect RT of these regions. For example, DDK concentration is higher close to centromeres, due to an interaction between DDK and the kinetochore complex[81], while Rif1 inhibits DDK close to telomeres by recruiting protein phosphatase 1 (PP1)[82]. Rif1 (Rap1 interacting factor 1) was initially identified in budding yeast as a protein involved in telomeric regulation and is highly conserved in eukaryotes[83]. It is one of the well described *trans* factors responsible for RT regulation, from budding yeast (described above) to humans[84]. Rif1 regulates RT in metazoans by playing a role in the organisation of high order chromatin structure within the nucleus: it positions some late replication domains in the nuclear periphery and constrains contacts between different replication domains[85]. The nuclear localisation of different replication domains will affect their ability to compete for limiting factors, and consequently, their timing of replication. Metazoan replication domains will be described in more details in subsequent sections.

Another example of *trans*-acting factors regulating the accessibility of limiting initiation proteins to origins includes the forkhead box transcription factors (TFs) Fkh1 and Fkh2, whose binding sites are enriched near some early origins and depleted from late origins in budding yeast[86]. These TFs are necessary for the clustering of a subset of early origins, thus providing an advantage in the competition for Cdc45[86]. In a similar situation to the early replication of budding yeast centromeres, it was shown that Fkh TFs can drive early origin firing by directly recruiting the limiting factor Dbf4 to a subset of origins[87]. The Fkh-regulated origins are the ones which fire early even when moved to different chromosomal positions (described above)[70]. Interestingly, disruption of the motifs involved in clustering of early origins affects the timing of these regions without affecting expression of surrounding genes[88], showing that Fkh1/Fkh2 regulate RT and transcription independently. Altogether these studies suggest that both the levels and origin's accessibility to limiting factors dictate the RT programme, and illustrate potential mechanisms that regulate origins ability to compete for them.

1.6.4 - The recycling model of DNA replication timing control

Interestingly, most limiting initiation factors do not remain bound to the active helicase during replication fork progression[89], suggesting that they are released from the replisome after origins have fired. A recent preprint from the Zegerman lab has shown that in budding yeast this release is actively regulated by phosphatases and is required for origin firing genome-wide[90], illustrating its biological importance. The fact that origin firing is dictated by the availability of limiting factors and ability of origins to compete for these factors, which are released once origins have fired, together with the observations from Mantiero et al. and others (described above) support the recycling model for RT control (Fig. 1.5). I have introduced this model in previous sections while describing the random gap problem. This is further supported by the fact that S-CDK and DDK activities are required throughout the full extent of S-phase[91,92]. As S-phase progresses and early origins fire there are less unfired origins relative to the pool of limiting initiation factors, increasing the probability of further origin firing, potentially explaining how origin firing efficiency increases as S-phase progresses[63], reducing the chance of DNA remaining un-replicated (Fig. 1.4 – random gap problem).

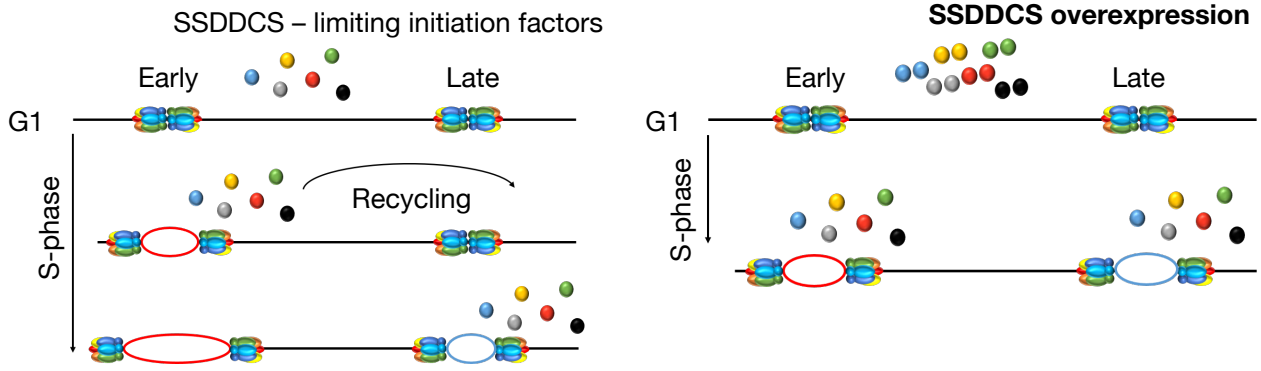


Figure 1.5 – The recycling model of DNA RT control. Left – Early and late origins are licensed with two opposing helicases in G1 phase, but not all origins fire at the same time during S-phase. Sld2, Sld3, Dpb11, Dbf4, Cdc45 and Sld7 (SSDDCS, represented as circles with different colours) are found in low concentration in cells. Early origins will fire at the start of S-phase, due to their increased affinity to these factors. Once early origins have fired, these factors are released and drive the firing of late origins, dictating the RT programme. **Right** - overexpression of SSDDCS in G1 bypasses the control of RT by recycling, as high levels of the initiation factors allow simultaneous firing of early and late origins[76,77].

These findings support the view that DNA RT is an actively regulated process which is influenced by the chromatin landscape and chromatin binding proteins. Several other cellular processes are regulated by the structural context of the genome and correlated with RT, such as transcription, histone mark deposition, the establishment of chromosomal domains and sub-nuclear chromatin arrangements[6]. Surprisingly, a complete understanding of the biological role of RT has remained elusive. During the next sections, I will discuss the links between RT and other genomic events, potentially providing possible reasons for a defined order of origin firing.

1.7 - The relationship of replication timing with transcription and chromatin organisation

The first observations that support a relationship between DNA RT, gene expression and chromatin were made more than 60 years ago, when Lima de Faria observed that transcriptionally active euchromatin is replicated before heterochromatin[3]. I have described how the chromatin architecture can affect origin firing, but the causal link between the two has remained elusive, which is also the case for RT and gene expression. Two opposing (but not mutually exclusive) models could explain the link between RT and transcription through differential chromatin architecture: 1) the higher accessibility of euchromatin makes it more permissive for both replication and transcription or 2) the RT of a genomic location affects its chromatin structure, and as such, its gene expression patterns.

Despite the link between RT and transcription, which suggest a mechanistic coordination and biological significance, many examples of regions where the two are not correlated have been observed in all organisms studied so far.

1.7.1 – Replication timing and gene expression are (broadly) correlated, but some exceptions exist

A long standing correlation in metazoans is the late replication of silenced heterochromatin[93]. Moreover, plasmids injected in cells at different times during S-phase have different transcriptional competence, i.e. plasmids injected early in S-

phase are better templates for transcription[94]. A treatment with trichostatin A (TSA), a histone deacetylase inhibitor, increased the expression of plasmids injected in late S-phase, suggesting that these are repressed due to the packaging of DNA with deacetylated histones[94]. However, this study used exogenous plasmids and as such does not show an association between RT and transcription of endogenous DNA. Initial studies have shown that gene-rich domains of open chromatin tend to replicate early, from humans[95,96] to mouse[97] and *Drosophila*[98] and many groups have shown a positive correlation between RT and probability of a gene being expressed[95]. Very recently, it was shown that perturbations in RT caused by loss of Rif1 (described in previous sections) are coupled with alterations in histone modifications and 3D chromatin compartments, but only have a limited effect on gene expression[99].

Considering that most mechanisms regulating DNA replication and the RT programme itself are conserved across eukaryotes, the overall lack of correlation between RT and the probability of transcription observed in budding yeast in the Raghuraman study came as a surprise[4]. I should mention that this correlation was addressed only for the 137 ribosomal protein genes, which account for ~50% of transcription from RNA polymerase II, tRNAs and small nucleolar RNA genes. Although the T_{rep} of these genes was not significantly different from the genome average, it should be mentioned that this study, despite being a key contributor to the field, was performed under one single growth condition, and the authors postulate that RT might change when cells adapt to different environmental conditions[4]. Omberg et al. have also found that ~88% of global mRNA expression in budding yeast is independent of DNA replication[100]. Finally, another study in budding yeast analysed the T_{rep} of cell cycle regulated genes and found that the 100 highest expressed genes tend to be earlier replicated while the 100 lowest expressed genes tend to be late replicated[101], but these differences are subtle.

Rhind and Gilbert suggested that the lack of correlation in budding yeast could be due to the differences in genome size compared to metazoans (100-fold smaller)[6]. The larger genomes of metazoans have a clear separation between early and late replication domains, which represent regions of the genome spanning many mega-bases which are replicated at similar times during S-phase, due to the coordinated

firing of origins. Initial studies from Hiratani et al. showed that during the first half of S-phase there is no correlation between RT and probability of transcription in human cell lines[7]. The correlation becomes stronger in mid to late S-phase, where earlier replication associates with transcription probability (more recent studies have dissected the nature of this correlation, which I will discuss in subsequent sections). Therefore, the budding yeast genome can be thought as the equivalent to a single early domain, with the exception of late replicated telomeric heterochromatin which represents a small portion of the genome[6]. The very short S-phase (~30 minutes compared to several hours in metazoans) could also affect the identification of meaningful correlations.

1.7.2 – Early replication of histones and gene dosage buffering

The histone genes represent an exception to the lack of correlation in budding yeast, as all 8 histone genes are highly expressed in S-phase and have a significantly earlier T_{rep} compared to the genome average[4]. The histone genes represent an interesting example of the relationship between RT and transcription, and a recent study has shown that early replication of these genes is required for their maximal expression[102]. Müller et al. have shown that the deletion of 3 origins proximal to a pair of histone genes, which significantly delayed the T_{rep} of this location, decreased their expression levels. More importantly, the T_{rep} of the rest of the genome was unaffected, and as such a pair of histone genes located on a different chromosome had no changes in expression[102]. Despite the down-regulation of the histone genes when they were replicated late, their expression was still timely induced during S-phase, suggesting that RT is important for the transcriptional rate but not for transcriptional activation of the histone genes.

High levels of histone mRNA are required during S-phase for the correct chromatin deposition in newly synthesised DNA[103], so their early replication ensures a tight coupling between expression and packaging of replicated DNA. This mechanism of increase in relative copy number associated with increased transcription is widely used by bacteria, as genes required for transcription and translation are located close to origins of replication and as such are present at higher doses[104]. In a set of

elegant experiments, Slager et al. have shown that genes involved in competence, which allow bacteria to take up foreign DNA from the environment, are also located close to origins. This organisation allows bacteria to increase the copy number of these genes upon antibiotic treatments targeting DNA replication, promptly activating competence and inducing resistance[105].

While this copy number effect is present in histone genes in budding yeast, Voichek et al. described a buffering mechanism mediated by the acetyltransferase Rtt109 that acetylates histone H3 at lysine 56 in newly replicated DNA and down-regulates gene expression during S-phase, maintaining expression homeostasis[106] (Fig. 1.6). Müller et al. analysed *rtt109Δ* gene expression data from Voichek et al. and found that histone genes are not buffered by this mechanism[102] (Fig. 1.6 - right). Interestingly, ~20% of yeast genes which are cell cycle regulated or induced by stress were excluded from Voichek analysis and the histone genes are part of this group[106]. This work suggests that expression of at least the 500 earliest replicating genes, which were compared against the 500 latest replicating, is buffered during S-phase, and histone genes represent one of the exceptions, explaining why their early replication is required for maximal expression.

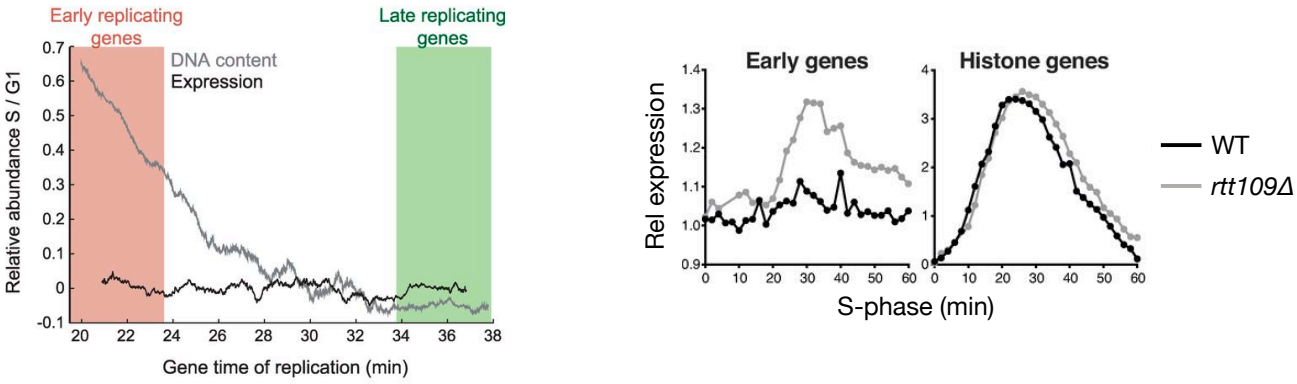


Figure 1.6 – Gene dosage compensation during S-phase in budding yeast. Left – Average expression of early replicating genes is buffered against changes in copy number, so that their expression levels are kept at levels similar to the average of late replicating genes which are yet to be replicated. Figure from Voichek et al[106]. Early and late replicated genes were defined as the 500 genes with lowest or highest replication timing, respectively. **Right -** Relative expression levels (S-phase time point over G1 arrested) determined by Voichek et al.[106] for early replicated genes and histone genes. Deletion of the acetyltransferase Rtt109, which buffers expression against changes in copy number, leads to the up-regulation of the early replicated genes, while expression of histone genes is not affected. This result shows that histone genes represent one of the exceptions to this buffering mechanism. Figure adapted from Müller et al.[102]

The authors suggested that this buffering mechanism ensures that expression homeostasis is maintained during S-phase so that genes which are early replicated, and as such face an increase in copy number in early S-phase, have similar levels of expression compared to late replicated genes (Fig. 1.6). It is still not clear whether such a buffering system exists in metazoans, but some studies suggest this is the case for a subset of genes[107].

The copy number effect affecting histone gene expression in budding yeast suggests a link between RT and transcriptional rate (i.e. gene is more or less expressed) rather than transcriptional activity (i.e. gene is expressed or silenced). However, most of the significant correlations described to date are between RT and transcriptional activity (i.e. probability of gene being expressed)[7,95]. Still, there are many examples of genes that are transcribed when late replicated and silenced when early replicated. Stress and apoptosis genes, for example, are early replicated despite the fact that they are silenced for most of the time, suggesting that RT could be important for transcriptional potential rather than transcription itself, as these genes need to be rapidly transcribed under certain environmental conditions[97]. As mentioned previously, the link between RT and transcription could just be a consequence of another correlated feature: early replication is correlated with open chromatin[96,108] (despite the existence of regions where the two are not correlated) and histone marks[109], some of which are associated with transcriptional activity[8]. Moreover, mathematical models were able to predict cell-type specific RT based on DNAse hypersensitivity profiles[110]. The interconnectedness of replication timing, transcription and chromatin makes the dissection of the exact relationship between the three a very challenging problem, as an effect in one will necessarily affect the others.

1.7.3 – Replication timing and the 3D organisation of the genome

The most striking association between any genomic feature and RT in metazoans is the 3D organisation of chromatin in the nucleus[111]. As described previously, metazoan genomes are organised in early and late replication domains separated by timing transition regions[6]. These domains and its boundaries align surprisingly well with topologically associating domains (TADs), suggesting that TADs act as the stable units of replication timing in metazoans[111,112] (Fig. 1.7).

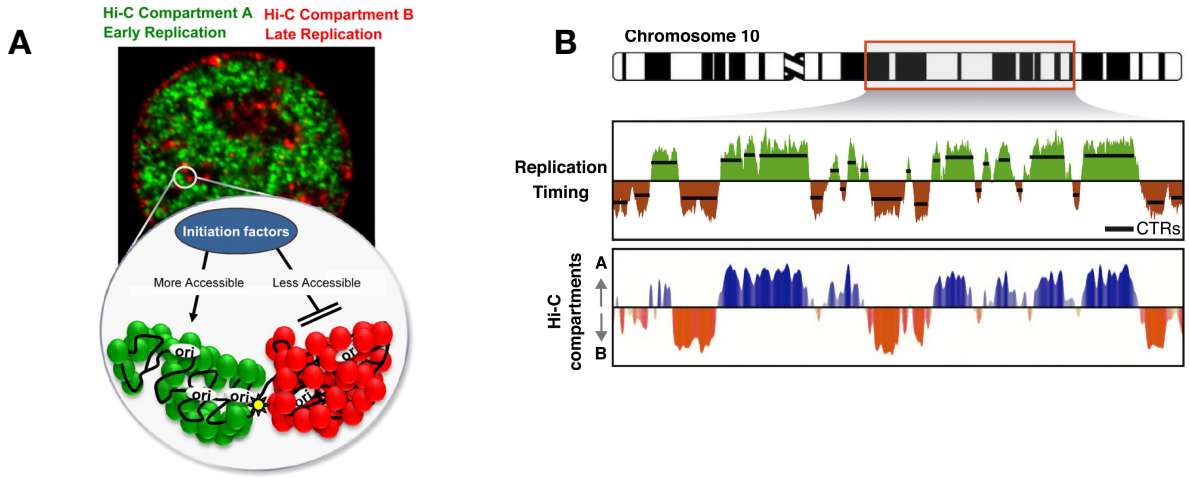


Figure 1.7 – Replication timing is associated with the 3D organisation of the nucleus. **A** – Cells were pulse labelled with two fluorescent nucleotide analogues in two different S-phase time-points (green and red during early and late S respectively) and imaged with specific antibodies for each. The A compartment correspond to early replicating domains which are located in the nuclear interior, while the late replicating B compartment is localised in the nuclear periphery and nucleolus (heterochromatin). The cartoon is a schematic view of a pair of adjacent early and late domains, which have different levels of accessibility for limiting initiation factors as described in previous sections. **B -** Alignment of replication timing and Hi-C data for a region of the human chromosome 10 confirms that the A compartment is associated with early replicating regions while the B compartment is associated with late replicating regions. CTR – constant timing regions. Figure modified from Rivera-Mulia et al.[113]

TADs were identified using chromatin conformation capture methods (Hi-C) and represent genomic regions in which DNA sequences exhibit significantly higher interaction frequency compared to sequences outsides the TAD[114] and are formed by a loop extrusion mechanism mediated by cohesin and CTCF proteins[115]. These interaction domains form two independent nuclear compartments: the A compartment corresponding to early replicating euchromatin localised in the nuclear interior and the B compartment corresponding to late replicating heterochromatin localised in the nuclear periphery and nucleolus[8] (Fig. 1.7).

TADs seem to be absent in budding yeast, which might again be a consequence of its small genome. Still, one study identified chromosome interaction domains in budding yeast, which are smaller compared to TADs from metazoans but have the same number of genes[116]. Using similar techniques, a recent work identified TADs in budding yeast as genomic regions regulating the synchronous firing of replication origins[117]. However, these domains identified as TADs do not share all characteristics with the mammalian counterparts, such as similar transcriptional activity within the same TAD. A recent study analysed chromatin 3D dynamics during the cell cycle in budding yeast and the authors observed that the read coverage of raw Hi-C libraries reflects the replication progression during S-phase[118]. These results support a link between structural organisation of the genome and DNA replication timing, from budding yeast to metazoans.

Despite the link between TADs and RT, there is evidence suggesting that these domains are not required for RT. RT patterns are present in one-cell mouse embryos[119], while TADs only form after the 4-cell embryo, in a replication dependent manner but independent of zygotic activation[120]. Further studies have shown an uncoupling between RT and genome organisation during embryogenesis, which I will discuss in the next section. TAD organisation is regulated by cohesin proteins and CTCF, and a recent study has shown that conditional degradation of cohesin does not affect replication patterns, both in asynchronous populations and in synchronised populations prior to entry in S-phase[121]. This study is in agreement with another study that showed that CTCF depletion has no major effects in RT genome-wide[122].

Interestingly, in this work the authors identified early replicating control elements (ERCEs) as *cis* elements that regulate CTCF-independent chromatin compartmentalisation. These elements do not overlap with the most efficient initiation zones and their deletion causes changes in RT, switch between A and B compartments and weakening of TAD architecture[122]. The authors started by analysing a single mouse TAD that becomes late replicated during loss of pluripotency, coincident with repression of genes located in this domain, movement to the nuclear periphery and chromatin rearrangements[34], and then identified regions across the genome that have a similar impact when deleted[122]. These

elements are occupied by proteins that promote local histone acetylation, so the authors suggest that they advance replication in a similar way to forkhead TFs in yeast[87], by increasing the ability of origins to compete for initiation factors[122]. This study has found that ERCEs also represent binding sites for transcription-factors, further supporting a role for TFs in RT regulation[122].

Despite all the studies reporting genome-wide correlations, it has been difficult to establish a causal link between RT and gene expression. Moreover, it is possible that RT is just a consequence of a defined structural organisation of chromatin in the nucleus. During the final section, I will discuss the links between RT and gene expression under different physiological contexts in order to shed some light into the biological function of a defined temporal pattern of origin firing.

1.8 - What is the biological role of the replication timing programme?

As I have described in previous sections, the temporal order of origin firing is conserved across eukaryotes, but its biological significance has remained a mystery. In order to understand the biological importance of the RT programme we must distinguish two separate concepts: why is S-phase longer than its minimum possible length and what is the reason for a defined temporal order of origin firing[6]? In the final section, I will distinguish these two effects (origin firing rates vs origin timing) and describe possible roles for this defined organisation of replication dynamics.

1.8.1 – Origin firing rates vs timing effects

As described in previous sections, some limiting initiation factors are present in lower levels compared to the total number of origins. As such, early origins which are more accessible to these factors fire before the less accessible late origins. Overexpression of these factors allows the early firing of late origins, which was demonstrated by Mantiero et al.[76] using the SSDDCS strain described in previous sections. In this study, the authors also observed that the increase in simultaneous origin firing in early S-phase caused by SSDDCS overexpression induced activation

of the Rad53 kinase. It is well described that the RT programme is one of the targets of the checkpoint Rad53, which blocks late origin firing under replicative stress conditions by inhibiting initiation factors such as Sld3 and Dbf4[123]. The authors hypothesised that depletion of dNTPs caused by simultaneous origin firing in early S-phase could potentially increase fork stalling events, resulting in the observed Rad53 activation[76]. This hypothesis was confirmed by artificially increasing the dNTP pool through the deletion of the ribonucleotide reductase Sml1, which suppressed Rad53 activation in the SSDDCS strain. Another study from the Zegerman lab used the SSDDCS strain to show that an increase in origin firing in early S-phase causes DNA topological stress[124], which could be explained by the fact that topoisomerases become limiting and cannot resolve all topological constraints caused by the simultaneous increase in the number of active replication forks. This work has also showed that another potential role of the RT is to avoid topological stress caused by collisions between the replication and transcription machinery.

There are examples of cases where a variation in the number of origins that fire during S-phase does not impacting the global RT programme. For example, progression into cellular senescence, a cell cycle arrest state which can be caused by exhaustion of proliferative potential (replicative senescence) or oncogene hyper-activation (oncogene induced senescence), causes replication stress through the slowing of fork rates and activation of dormant origins, but has no significant effect on the RT programme[125]. Oncogene-induced replication stress can cause both origin over and under-usage[126], suggesting that RT is highly robust against replicative stress and variations in origin firing rates. Moreover, a study in yeast has shown that mutations in genes involved in cell cycle control, replication machinery and dNTP synthesis have an extended S-phase due to a delay in origin fire, but this delay was proportional to the S-phase duration, meaning that the relative order of firing was kept[127]. These studies suggest that cells regulate the number of origins that fire during S-phase to ensure that limiting factors involved in various cellular events are not exhausted, and that RT is robust against changes in origin firing rates.

1.8.2 – Replication timing is regulated during development

The fact that the RT is fairly conserved within mammalian cell types[128], but roughly 50% of the genome changes RT during cellular differentiation[34], suggests that RT could be exploited as a mechanism to induce cell fate transitions. Importantly, not just DNA but all epigenetic information has to be maintained, so a cell-type specific RT could ensure the transmission of cell-type specific epigenetic states, in addition to a mechanism to induce changes in these states during cell fate transitions[34]. The RT changes observed during cellular differentiation occur in units of 400-800 kb corresponding to the replication domains described in previous sections, and are usually associated with changes in transcriptional regulation for a certain class of genes[129].

This model suggests a positive feedback loop mechanism (i.e. RT affects chromatin and chromatin affects RT), as a shift in RT in a given location is going to affect the RT of surrounding regions by passive replication, which would modify the chromatin landscape and as such reprogramme the RT of whole genomic regions[6], which would reconcile epigenetic inheritance and developmental reprogramming.

Consistent with this idea, genome-wide studies have found that domains that change their timing from early to late during differentiation maintain their late replication in differentiated cells, suggesting that early replication represents a "pluripotency fingerprint" and late replication could act as an epigenetic barrier to their reprogramming back to stem cells[34]. The development of stem cell *in vitro* systems allowed the analysis of the relationship between RT and transcription in the context of differentiation, and have shown that the two are coordinated[7,34,129].

Recent studies have demonstrated that early constitutive genes (genes that are early replicated in every cell type) seem to drive the positive correlation between RT and transcription, and these tend to be also highly expressed. On the other hand, the majority of genes that change RT during differentiation are expressed when late replicated[129]. Moreover, in human stem cells undergoing differentiation, many RT changes are independent from gene expression changes, despite the global correlation between the two[130]. These findings suggest that despite their close

association, RT and transcription could be independently regulated during differentiation.

A recent study might help reconcile this lack of association. Rivera-Mulia et al. analysed RT profiles and transcriptomes from 15 human cells lines undergoing differentiation in order to identify RT regulatory networks[131]. The authors have found that the expression patterns of key TFs involved in cellular differentiation were correlated with the RT of downstream differentiation regulator genes. Interestingly, these TFs tend to bind sites with affected T_{rep} during differentiation[131]. This observation is in agreement with the study that has found ERCEs as binding sites for various TFs[122] and could help explain the examples of lack of correlation between RT and gene expression, suggesting a mechanism in which TFs regulate RT independently of their role in gene expression regulation[131]. This is consistent with the Fkh1/Fkh2 TFs in budding yeast, which regulate RT and transcription independently[86]: disruption of the motifs involved in clustering of early origins affects the timing of these regions without affecting expression of surrounding genes[88]. All these findings illustrate the complex nature of the relationship between RT and gene expression.

Being a simple unicellular eukaryotic organism, budding yeast does not have complex cellular differentiation events comparable to metazoans. Still, there are examples of differential origin usage during pre-meiotic S-phase, an event that precedes spore formation and is triggered under poor nutrient conditions[132]. The most efficient origin used during mitotic S-phase is inhibited in pre-meiotic S-phase. This origin is located in the open reading frame of a gene that is transcribed during early stages of meiosis and expression of this gene coincides with suppression of initiation in this origin, even when the gene was conditionally overexpressed during mitosis[133]. Another study has identified mitosis and meiosis specific Mcm2-7 binding events, and while the mitosis specific binding events were found in sporulation-induced genes, the meiosis specific binding events were found in mitotic budding-related genes[134]. In this particular context, the data suggests that origin usage is directly affected by transcription, through the removal of pre-RC components from origins located within open reading frames of genes expressed in mitosis or meiosis[133,134]. These mitotic and meiotic-specific sites represent about

10% of all Mcm2-7 binding sites identified, and illustrate an example of coordination between replication and transcription associated to changes in cellular physiology in budding yeast.

1.8.3 – Replication timing and monoallelic expression

RT is also associated with monoallelic expression. The best known case is the inactivation of one of the female X chromosomes in mammals, which is accompanied by a switch to late replication[135]. The β-globin gene represents an example of both monoallelic expression and developmental control of RT: in erythroid cells this locus is early-replicated, highly acetylated and transcribed, while in non-erythroid cells it is late-replicated, not acetylated and silenced[136]. Finally, the rDNA gene cluster which consists of several copies of ribosomal genes also follows similar regulation: late replicated and highly methylated copies are silenced while early replicated unmethylated copies are expressed[137].

Moreover, Koren et al. have identified base-pair differences between individuals that cause differences in the timing of replication of those locations (replication timing quantitative trait loci or rtQTLs[138]). These rtQTLs were associated with the differential usage of origins of replication and with gene expression variation at mega-base scales. More recently, the same authors have expanded this work to human pluripotent cell lines and have identified 1617 rtQTLs, which were also associated with particular histone modifications and pluripotency-related TFs[139].

1.8.4 – Modulation of replication timing affects zygotic transcription

As I mentioned in the previous section, RT patterns are present in embryos in the absence of TADs or zygotic transcription[119]. Metazoan embryos undergo rapid divisions and are largely transcriptionally silent. The maternal to zygote transition consists of degradation of maternal mRNAs, activation of zygotic transcription and lengthening of the cell cycle[140]. In the rapid dividing embryos of *Drosophila*[141], zebrafish[142] and *C. elegans*[143] RT patterns precede zygotic activation. A study from the Zegerman lab has shown that perturbations of RT induced by overexpression of limiting factors in *Xenopus* embryos, which increases the number of initiation

events, affect the mid-blastula transition and the start of zygotic transcription[144]. This data supports a model in which the lengthening of the cell cycle associated with zygotic transcription is important for embryonic development, further illustrating roles of RT in the regulation of gene expression during differentiation.

1.8.5 – Replication timing and mutation rates

Finally, the RT programme is also associated with mutation rates: it has been shown that regions of the genome which are late replicated tend to have increased mutation rates[145]. There is evidence supporting the view that repair mechanisms are less efficient in late replicating heterochromatin, which would drive evolutionarily pressure to replicate gene-rich regions and house-keeping genes early in S-phase[146]. As such, RT would concentrate genomic variation to defined regions of the genome, which agrees with observations that genes involved in speciation events tend to be late replicated and located in mutational hotspots[146,147].

One study has shown that this is also the case in budding yeast, as changing the chromosomic location of the *URA3* gene affected its mutation rate in a replication timing dependent manner: the mutation rate was higher when the gene was moved to late replicated regions, compared to early ones[148]. Moreover, delaying RT by removing replication origins increased the mutation rate of these regions, further supporting that late replicating regions are mutational hotspots[148]. The fact that essential genes in budding yeast tend to be significantly closer to early replicating centromeres[149], suggests evolutionary pressure to keep these genes in regions of low mutation rates.

1.8.6 – Replication timing and cancer

While mutations are an important part of evolution and adaptive radiation events, they can also be harmful for organisms and cause diseases such as cancer. There are plenty of studies showing an association between mutational rates and RT in cancer[147,150] and several other diseases[151]. A genome-wide study of RT in paediatric leukaemia tumours, has shown that replication profiles from normal B and T cells were different from tumours, and despite the heterogeneity between tumour

samples, a known leukemic translocation site was affected in all tumours[152]. Moreover, there are several examples of cancer-specific genes with changes in RT[147] and one large scale genome-wide study has identified both early and late cancer domains common to several cancer types[153].

Related to the rtQTLs described in the "Replication timing and monoallelic expression" sub-section, Koren et al. identified one rtQTL at the *JAK2* locus which was associated with increased mutation rates in this locus, most likely due to the increased collisions between the replication and transcriptional machinery caused by the early replication of the *JAK2* nearby origin[138]. These mutations are associated with increased *JAK2* expression that result in myeloproliferative neoplasms[138], illustrating how differences in RT could affect the mutation rates of medically relevant alleles.

The role of RT on gene expression, mutation rates, chromatin architecture and overall genome fitness, further supports its key role in carcinogenesis. However, the close association between these cellular processes makes it very hard to determine the exact sequence of events and more studies will be required to dissect these relationships. A recent preprint has shown that low replication stress has a stronger effect on RT of cancer cells compared to non-cancer cells, and that these changes affect gene expression and chromatin remodelling and are transmitted to daughter cells, suggesting a mechanism used by cancer cells to adapt to environmental stress[154]. As such, targeting RT regulators in combination with chemotherapies could be used as a strategy to supress cancer cell resistance to replicative stress.

1.9 - Work presented in this thesis

Many studies have analysed the biological role of RT and its relationship with gene expression and chromatin architecture, illustrating the complex mechanistic links between them. A complete understanding of how these processes are regulated is still missing, but the advance of genomics together with manipulation of the RT programme is starting to reveal some causal links. It is clear that RT, gene expression and chromatin affect each other under several different cellular contexts, from differentiation to disease, and the conservation of RT across organisms

suggests an essential biological role, which is still not fully understood. During my thesis I used a conditional system to perturb DNA replication timing in a single cell cycle in budding yeast, in combination with whole-genome sequencing techniques such as replication profiles, RNA-Seq and MNase-Seq in order to address the impact of a perturbed RT in the genome function and structure. Overall, dramatic changes in gene expression, chromatin structure and TF binding events were observed. Some, but not all differentially expressed genes showed significant chromatin changes. Conversely, some genes with changes in the chromatin landscape which showed no changes in expression were also identified, further supporting the complex nature of the relationship between RT, gene expression and chromatin. Differential TF binding events explained some of the changes observed, supporting a role for RT to maintain the correct TF binding dynamics during S-phase. Additionally, the fact that budding yeast origins are defined by specific sequences allowed the local modulation of RT and analysis of the direct effect of RT on gene expression. Altogether, the work generated during this thesis provides insight into the complex relationship between replication timing, gene expression and the chromatin landscape.

Chapter 2 - Methods

2.1 - Yeast-related methods

2.1.1 - Yeast strains used in this study

Strain	Relevant genotype
PZ356	*W303a MATa ade2-1 ura3-1 his3-11, 15 trp1-1 leu2-3, 112 can1-100 rad5-535 sml1Δ::URA3*
PZ523	*W303a MATa ade2-1 ura3-1 his3-11, 15 trp1-1 leu2-3, 112 can1-100 rad5-535 leu2::Sld7-PGAL1-10-Cdc45::LEU2 his3::SLD3-A-PGAL1-10-Dbf4-A::HIS3 trp1::Sld2-PGAL1-10-Dpb11::TRP1 sml1Δ::HphNT*
PZ1407	*W303a MATa ade2-1 ura3-1 his3-11, 15 trp1-1 leu2-3, 112 can1-100 rad5-535 leu2::Sld7-PGAL1-10-Cdc45::LEU2 his3::SLD3-A-PGAL1-10-Dbf4-A::HIS3 trp1::Sld2-PGAL1-10-Dpb11::TRP1 sml1Δ::HphNT ARS1008Δ ARS1009Δ*
PZ1435	*W303a MATa ade2-1 ura3-1 his3-11, 15 trp1-1 leu2-3, 112 can1-100 rad5-535 sml1Δ::URA3 ARS1008Δ ARS1009Δ*
PZ3004	*W303a MATa ade2-1 ura3-1 his3-11, 15 trp1-1 leu2-3, 112 can1-100 rad5-535 sml1Δ::URA3 ARS816Δ ARS818Δ::KanMX*
PZ3005	*W303a MATa ade2-1 ura3-1 his3-11, 15 trp1-1 leu2-3, 112 can1-100 rad5-535 leu2::Sld7-PGAL1-10-Cdc45::LEU2 his3::SLD3-A-PGAL1-10-Dbf4-A::HIS3 trp1::Sld2-PGAL1-10-Dpb11::TRP1 sml1Δ::HphNT ARS816Δ ARS818Δ::KanMX*

2.1.2 - Yeast media

The medium used to grow yeast was YP medium, autoclaved prior to use and supplemented with 2% raffinose unless stated to the contrary in the appropriate text. In order to select for marker genes, such as *TRP1*, *HIS3*, *LEU2* or *URA3*, minimal medium was used, without the relevant amino acid. Saturated cultures were mixed with 15% glycerol prior to long-term storage at -80°C where necessary. Yeast plates were maintained at 4°C for short-term storage.

2.1.3 - Block and release time-course

The *sml1Δ* and *sml1Δ SSDDCS S. cerevisiae* strains were grown overnight in YP-raffinose at room temperature. After ensuring that cultures were growing exponentially, 100ml was transferred to 30°C shaking water bath for one cell cycle. At $1x10^7$ cells/ml, 90µl of stock solution of alpha factor was added to 100ml of culture (1:900 dilution) and after 90 minutes 45µl of stock solution of alpha factor (5mg/ml) was added. To confirm the G1 arrest, cells were analysed under microscope using a haemocytometer and the arrest was considered successful if more than 95% of cells had the G1 characteristic shape ("shmoo") or were unbudded. Upon arrest, 10ml of 20% galactose was added to the cultures to induce the overexpression of the six factors. 30 minutes post galactose addition, G1 samples were collected. Then cultures were washed twice with fresh YP-galactose to release cells from G1 arrest and resuspended in 100ml of YP-galactose. Cultures were maintained at the 30°C shaking water bath for 60 minutes, and 8ml was taken every 5 minutes for MNase-Seq, as well as 500µl for FACS.

2.1.4 – Collection of samples for whole-genome sequencing (replication profiles)

Yeast genomic DNA was extracted using the smash and grab method (https://fangman-brewer.genetics.washington.edu/smash-n-grab.html). DNA was sonicated using the Bioruptor Pico sonicator (Diagenode), and the libraries were prepared according to the TruSeq Nano sample preparation guide from Illumina.

2.1.5 – Collection of samples for whole-transcriptome sequencing (RNA-Seq)

Yeast RNA was extracted using the Ambion RiboPure – Yeast Kit.

2.1.6 - Digestion of chromatin with MNase for mono-nucleosome analysis (adapted from Nocetti and Whitehouse, 2016)[155]

Day 1 - 8ml of yeast culture collected in each time-point was centrifuged for 2 minutes at 4000 rpm and resuspended in 40ml of 1x PBS, 1% formaldehyde. Samples were mixed and left shaking gently for 10 minutes at room temperature on gyro-rocker to crosslink DNA and proteins. Crosslinking was quenched by adding 5ml of 2.5M glycine and left shaking for 10 mins on orbital shaker at room temperature. Samples were left on ice until all samples have been collected. Then, samples were spun for 5 min at 3200 rpm in 50 ml tubes and washed with 50 ml of sterile ddH_2O. Pellets were resuspended and transferred to 1.7ml Axygen tubes, and spun down at top speed in table top centrifuge. Liquid was carefully aspirated and pellets vortexed. Pellets were resuspended in 950µL of zymolyase digestion buffer (ZDB: 50 mM Tris Cl at pH 7.5, 1 M sorbitol, 10 mM β-mercaptoethanol) to remove the cell wall. Then, 100 µL of freshly prepared zymolyase solution (10mg/ml dissolved in ZDB) was added to each sample, and digestion was performed for 60 min at 30°C shaking gently in water bath. Efficiency of digestion was assessed by checking cell morphology under the microscope: cells with digested cell walls will appear spherical. A second test is to take 1µl and dilute to 20µl with ddH_20. As cells no longer have a cell wall the osmotic shock will burst them. So absence of cells means zymolyase treatment was successful (spheroplasting). Spheroplasts were pelleted in a microfuge, at 5000 rpm for 5 minutes at 4°C and washed with 1 ml of ZDB. Pellets were then resuspended in 1 ml of spheroplast digestion buffer (SDB: 1 M sorbitol, 50 mM NaCl, 10 mM Tris at pH 8, 5 mM $MgCl_2$, 1 mM $CaCl_2$, 1 mM β-mercaptoethanol, 0.15% NP40). Samples were spun down in the microcentrifuge and gently resuspended in 0.5 ml of SDB. Then 90U of MNase (9µl of 10U/µl MNase solution) was added, and tubes were well mixed and left with gentle agitation for 3 min at 37°C. The amount of MNase needs to be experimentally determined by titration with every batch of MNase. MNase digestion was stopped with the addition of 50 µl of 0.5 M EGTA. Tubes were vortexed after adding EGTA. Samples were treated with RNAse by adding 2 µl of RNase I (100units/µL) for at least 1 hour at 37°C. 10 µl of freshly made up stock of 10mg/ml Proteinase K was added and

samples were left at 42°C for at least 3 hours. The formaldehyde cross-links were reversed by incubating samples for > 6 h at 65°C.

Day 2 - samples were transferred to 2ml rubber-sealed screw-cap tubes. 1 volume of Phenol-Chloroform pH 8 (~570µl) was added to samples and vortexed and spun for 5 minutes. Aqueous phases were collected to new 2ml lo-bind Eppendorf tube, 5 µl of glycogen and 190 µL of 3M sodium acetate were added and samples were ethanol-precipitated with 1250 µL of cold absolute ethanol (2.5x) and vortexed. Samples were incubated at -20°C overnight and centrifuged at 13000 rpm for 30 minutes at 4°C.

Day 3 - Pellets were gently washed by adding 1ml of freshly made 70% ethanol. Then pellets were pulsed down quickly and most volume was carefully aspirated, the wash was repeated with 1ml of 70% ethanol and then tubes were spun down for 15 min at 4°C. Ethanol was carefully aspirated and pellets were air dried for 15 minutes at room temperature, 100µl of Illumina Resuspension Buffer was added and samples incubated for 1 hour at 37°C to redissolve DNA. Size range and relative molarity were determined on a Tapestation using D1000 and Genomic screen tape and total yield was quantified using Qubit broad range dsDNA kit.

2.1.7 - Digestion of chromatin with MNase for transcription factor binding analysis

Same protocol as the one used for mono-nucleosome analysis described above, with the following modifications in the MNase step: after spheroplasting and resuspension in SDB, 5U of MNase (5µl of 1U/µl MNase solution) was added and tubes were incubated on the bench (room temperature) for 20 minutes.

2.1.8 - Flow cytometry of yeast with Sodium Citrate buffer

500µl of yeast culture was spun down, then fixed in 500µl of cold 70% ethanol for 2 hours at room temperature (~20°C) or overnight at 4°C. After fixation, cells were centrifuged at 13300 rpm for 2 minutes and the pellet was washed with 1ml of 50 mM sodium citrate. Cells were then centrifuged and resuspended in 1ml of 50 mM sodium citrate with 10 µg/ml of RNase and incubated at 37°C for 4 hours. After the

RNase treatment, cells were centrifuged and resuspended in 50 mM HCl with 5 mg/ml of pepsin and incubated at 37°C for 30 minutes. Cells were then washed with 1ml of 50 mM sodium citrate. After centrifugation, cells were resuspended in 1ml of 50 mM Tris pH 7.4 with 0.5 µg/ml of propidium iodide (PI). Finally, tubes were vortexed and 100µl was added to FACS tubes with 1ml of 50 mM Tris pH 7.4 with 0.5 µg/ml of PI. Before processing in the cytometer, cells were sonicated for 8 seconds at 40% amplitude.

2.1.9 - Mating and tetrad dissection

To produce new combinations of genes, relevant MATα and MATa strains were crossed by mixing two cultures on a non-selective plate and incubated for at least 4h to create a new diploid strain. Diploid cells were isolated under a tetrad dissection microscope and grown on rich sporulation medium (RSM) for 3 or more days. After confirming the presence of tetrads under the microscope, tetrads were digested using 2µl of lyticase solution in 100µl of water for 2 minutes at room temperature. These tetrads were then dissected under a tetrad dissection microscope, and their genotype determined through PCR and marker selection. Mating type was determined by crossing the new strains with the tester strains DC14 and DC17 and replica plating onto minimal medium.

2.1.10 - Yeast transformation

10ml of $1x10^7$ cells/ml mid-log phase yeast cells were washed with ddH_2O, then resuspended in 1ml of buffer 1 (0.8ml H_2O, 0.1ml 10x TE pH 7.5 and 0.1ml 1M lithium acetate pH 7.5). Cells were then centrifuged at high speed for 5 seconds and most supernatant was removed, except for 50µl which was used to resuspend the pellets. Then, 5µl of freshly boiled and rapidly cooled salmon sperm ssDNA (10mg/ml) was added together with 1µg of the transformation DNA. Then 300µl of buffer 2 was added (0.8ml 50% PEG4000, 0.1ml 10x TE pH 7.5 and 0.1ml 1M lithium acetate pH 7.5) and the mixture was incubated for 30 minutes at 30°C. 100% DMSO was added to the final concentration of 10% and samples were heat shocked at 42° for 15 minutes followed by cooling on ice. Samples were then

centrifuged and resuspended in 0.5ml ddH2O and plated on the appropriate plate. If aminoglycoside antibiotics were utilised as selective markers, cells were grown in YPD for at least 3h prior to plating.

2.1.11 - Yeast genomic DNA extraction

10ml of yeast culture at $1x10^7$ cells/ml was centrifuged at 3200rpm for 2 minutes in screw cap rubber sealed tubes. Pellets were resuspended in 200µl lysis solution (10mM Tris pH 8, 1mM EDTA, 100mM NaCl, 1% SDS, 2% Triton X-100), 200µl of phenol/chloroform pH 8 (1:1) and 200µl of glass beads (0.45mm diameter). Tubes were vortexed for 30 seconds prior to the addition of 200µl TE, then vortexed again for 30 seconds. Cells were then centrifuged for 2 minutes at room temperature at full speed, and the 380µl of the aqueous layer was transferred to new Eppendorf tubes. 2 volumes of 100% ethanol were added, and samples were mixed by inverting the tubes a few times. Samples were then centrifuged for 2 minutes at full speed. After discarding the supernatant, the pellet was washed with 1ml cold 70% ethanol and briefly centrifuged. Then pellets were air dried at room temperature and resuspended in 50µl TE buffer (10mM Tris, 1mM EDTA pH 8) containing 50µg/ml RNase A. Samples were then incubated at 37°C for 1h to degrade RNA.

2.2 - Molecular biology

2.2.1 - Polymerase chain reaction (PCR)

PCR was performed using the Phusion High Fidelity DNA polymerase (NEB) in a 50µl reaction mixture of 5x Phusion HF Buffer, 1µl 10mM dNTP mixture, 2.5µl of both the forward and reverse primers diluted to 10µM, and template DNA. Reactions were then carried out using a peqSTAR 96x Universal gradient apparatus (PEQLAB), following the standard protocol: 98°C, 30 seconds; 98°C, 10 seconds, primer-dependent annealing temperature, 30 seconds, 72°C, 45sec/kb of product (35 cycles); 72°C, 5 minutes, final elongation. PCR products were visualised

following agarose gel electrophoresis and, when necessary, were purified using the QIAquick PCR purification kit (Qiagen) or the QIAquick gel extraction kit (Qiagen).

2.2.2 - Agarose gel electrophoresis

1% agarose gels were made using 1x TAE (40mM Tris, 20mM acetic acid, 1mM EDTA) containing 1µg/ml Sybr Safe. Samples were diluted in 6x loading buffer (0.03% w/v bromophenol blue, 60% glycerol, 10mM Tris pH 8, 60mM EDTA pH 8). Electrophoresis apparatus was run at 80-100V, and DNA bands were then visualised using a UV transilluminator, with their size estimated against a DNA ladder.

2.2.3 - Sanger sequencing

Sanger sequencing was carried out in the DNA sequencing facility of the Department of Biochemistry, University of Cambridge.

2.3 - Next-generation sequencing

2.3.1 - Library preparation – MNase-Seq mono-nucleosome reads

250ng of MNase-digested DNA from each sample was end-repaired using the Illumina TruSeq DNA nano kit. AMPure XP beads were added (1.8x volume of DNA, DNA >100bp on beads, <100bp in supernatant) to each reaction to purify the mononucleosomal fragments. A-tailing and adapter ligation was performed using the Illumina TruSeq DNA nano kit. Two subsequent steps of beads purification (1.4x volume of DNA) were performed in order to remove adapter dimers. Based on tests using hyperladder V (25bp bands) the beads can selectively retain DNA of ~270bp (mono-nucleosomal + adapters) from free adapters (60-120bp) if used at a 1.4x ratio to the volume of DNA sample. PCR cycle quantitation was performed for each sample using KAPA Syber Fast reagents and libraries were PCR amplified using the Illumina TruSeq DNA nano kit, followed by another step of bead purification (1.4x volume of DNA). Library quality and quantity were validated on a Tapestation using D1000 screen tape, Qubit broad range dsDNA kit and NEBNext library quantitation

kit for Illumina. Libraries were pooled to final 100nM molarity and one step of bead purification (1x volume of DNA) was performed to completely remove adapter dimers. Finally, 20µl of the pooled libraries at a final 20nM molarity was sequenced in a Illumina HiSeq 1500 platform by the Gurdon Institute Core NGS sequencing facility using 50 bp paired-end reads.

2.3.2 - Library preparation – subMNase-Seq transcription-factor reads

Same as previous with the following modifications: after end-repair and before A-tailing, samples were cleaned by performing a phenol-chloroform precipitation, in order to reduce the volume of the samples without using AMPure XP beads. For adapter ligation, 20% of the amount recommended by Illumina was used (adapters were diluted 1:10 in RSB), in order to minimize adapter dimer formation. This was done because adapter dimers cannot be removed using AMPure XP beads as their size is very similar to TF binding fragments, and smaller fragments are preferentially amplified during library preparation, which means we would be wasting sequencing depth with adapters. After ligation, samples were cleaned using 1.8x AMPure XP beads (DNA >100bp on beads), so TF binding events corresponding to 10-80bp footprint + two adapters (120bp) will bind to the beads.

2.4 - Bioinformatic analysis

Note: Code is written in `Monaco font size 10`. All bioinformatic software used are installed in the Gurdon Institute Bioinformatics cluster and were run in the cluster - Ubuntu 16.04.5 LTS (GNU/Linux 4.4.0-127-generic x86_64). Downstream analysis and figures were generated in Rstudio, both in my local machine (Version 1.0.136) or the Gurdon Rstudio remote cluster (Version 1.4.17).

2.4.1 - Quality control

All samples' quality was assessed using FastQC High Throughput Sequence QC Report version 0.11.4.

2.4.2 - Mapping

All samples were mapped using bowtie2 (version 2.2.6) to the budding yeast reference genome (strain S288C, version R64-2-1), which was indexed using `bowtie2-build`. SAM files were then converted to BAM, sorted and indexed using `samtools` (version 0.1.19). The quality control of the alignments was assessed using Qualimap (version 2.2.1).

2.4.3 - Replication profiles

Before generating the replication profiles, sequencing depth was normalised for each timepoint using a bulk value derived from the fraction of the genome that has been replicated at that timepoint (a value between 1 and 2). These values were derived using fitSigmoid https://dzmitry.shinyapps.io/flowfit/.

To generate replication timing profiles, the ratio of uniquely mapped reads in the replicating samples to the non-replicating samples was calculated following Batrakou et al[29]. Then, this ratio was plotted for each time-point, and a sigmoid line was fitted. T_{rep} was determined as the time of half-maximal replication (ratio = 1.5). Replication profiles were generated by plotting T_{rep} values for each chromosome location using ggplot2 and smoothed using a moving average in R. All downstream analysis was performed in R.

2.4.4 - RNA-Seq analysis

Read counts for each gene were extracted using genomic ranges and differential expression analysis was performed using DESeq2[156]. Pair-wise analysis of each time-point was performed using the Wald test. For the time-course analysis, the likelihood ratio test (LRT) was used. Genes were considered differentially expressed if the p-value adjusted value from these tests was < 0.01. PCA analysis was

performed using the variance stabilisation transformation (`vst` command from DESeq2) and plotted using ggplot2. For the k-means clustering, the gene expression data was normalised by row using the `scale` command and 6 clusters were generated using the `k-means` command with a maximum of 50 iterations. Heatmaps were generated in R using `heatmap.2`. Distance to origins, centromeres and telomeres was calculated using HOMER (v4.10.1)[157]. Statistical analyses were performed using R as described in the main text. Gene ontology enrichment analysis was performed using YeastMine[158] and visualised in R using REVIGO.

2.4.5 – Analysis of *rtt109Δ* RNA-Seq data

Gene expression data from Voichek et al.[106] was kindly provided by Dr. Yoav Voichek. Average transcript levels from cluster 1 genes were calculated as follows: relative expression was calculated for each gene in each time-point by dividing the signal in the relevant time-point by the signal of the same gene in G1 (alpha-factor synchronised). These ratios were $\log_2$ normalised, and the average relative expression of all cluster 1 genes was calculated for each time-point on each strain. *NDT80* was below the detection threshold in most samples, including the G1 time-point in *rtt109Δ* so the raw expression levels were plotted instead.

2.4.6 - Mono-nucleosome MNase-Seq analysis

Nucleosome calls were identified by processing the BAM files using DANPOS[159] (version 2.2.2): `danpos dpos` was used to generate wig files, perform the time-point pairwise analysis and identify the different classes of dynamic nucleosomes. `danpos profile` was used to generate the files required for the nucleosome profiles, which were plotted in R. The files with the genomic coordinates of the locations where the heatmaps should be centred were generated using USCS Genome Browser http://genome.ucsc.edu/cgi-bin/hgTables. Peaks in these profiles represent nucleosome dyads and valleys linker DNA or nucleosome depleted regions.

The +1 nucleosome was identified by calculating the distance of nucleosomes to promoters using HOMER. Nucleosomes within -20bp to 80bp of the TSS were classified as +1. The +1 relative position to the TSS was calculated by subtracting

the genomic position of the nucleosome to the TSS of the corresponding gene. ACF analysis was performed following Gutiérrez et al[160]. The autocorrelation function was used to determine the pattern of organisation of the first four nucleosomes (+1, +2, +3 and +4) within each gene, using the nucleosome sized reads between 140bp and 180bp overlapping each gene.

2.4.7 - Transcription-factor enrichment analysis

To identify TF enriched for the binding of different groups of genes, we used the rank sum test from YeTFaSCo[161] to compare different lists of genes with datasets of ChIP-chip. The two outputs from the rank sum test are a p-value representing how significant the association is and the area under the receiving operator characteristic (ROC) curve, which represents whether the list of genes provided is significantly enriched (ROC > 0.5) or depleted (ROC < 0.5) for targets which are bound by each TF present on the database. The results were plotted as a volcano plot using R.

2.4.8 - Identification of TF binding motifs genome-wide

To identify TF binding regions genome-wide, the sequence motif was extracted in meme format from the JASPAR database[162], which was used as an input to FIMO[163] to identify all genomic locations where this motif is present. These regions were annotated to genes using HOMER.

2.4.9 - Sub-nucleosomal MNase-Seq analysis

BAM files were converted into bigWig files for visualisation of the results using IGV. For this purpose, `bamCoverage` (version 3.0.2) was used with the following argument: `--binSize 1 --minFragmentLenght 0 –maxFragmentLenght 100`. Heatmaps of read coverage was generated by using the bigWig files as inputs to `computeMatrix` and `plotHeatmap` (version 3.0.2)

Identification of high confidence sub-nucleosomal peaks and calculation of fold-change differences between the strains was performed following Gutiérrez et al.[160]:

DNA fragments with less than 100bp (sub-nucleosomal events) were selected and the same number of reads was sampled for each fragment size using the time-point with the minimum number of reads for that fragment size. Then, all samples were merged to call all possible peaks in the data. A peak was considered high confidence if the sum of reads mapping to that peak (log_2 normalised) across all samples was higher than 75. The log_2 ratio of normalised reads occupying each peak between the two strains was calculated for each time-point. Heatmaps of peak-fold change were generated in R using `heatmap.2`.

Sub-nucleosomal peaks were annotated to TSS and TF binding sites using HOMER as described in section 2.4.6 and 2.4.8.

Chapter 3 – Effect of a perturbed replication timing programme on gene expression

3.1 - Overexpression of limiting initiation factors advances replication timing genome-wide

In order to perturb RT genome-wide, the conditional system developed in the Zegerman lab that allows the overexpression of 6 limiting factors under the control of a galactose inducible promoter[76] was used. Sld2, Sld3, Dpb11, Dbf4, Cdc45 and Sld7 (SSDDCS) are found in low concentration in budding yeast cells and as such are rate limiting for DNA replication initiation (Fig. 1.5). As described in the Introduction, dNTPs become limiting when these six factors are overexpressed[76], so the ribonucleotide reductase inhibitor *SML1* was also deleted in order to increase the dNTP pool and avoid Rad53 activation. As such, the six factor overexpression strain (*sml1Δ SSDDCS*) was compared to a control strain where *SML1* was also deleted (*sml1Δ*). The fact that budding yeast can be synchronised in G1 using the mating pheromone alpha-factor allows the following block and release experimental set-up: SSDDCS overexpression was induced in G1 arrested cells for 30 min, then cells were released into a synchronous S-phase and samples were collected every 5 minutes to analyse replication progression by FACS and whole-genome DNA sequencing (Fig. 3.1). See Methods for detailed protocols.

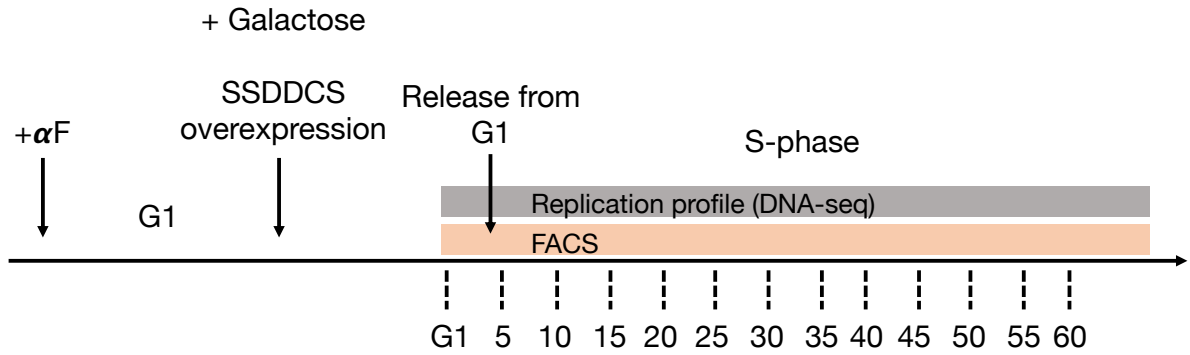


Figure 3.1 – Experimental set-up. – A population of budding yeast cells was arrested in G1 phase using the mating pheromone alpha-factor (+αF) and the overexpression of SSDDCS was induced for 30 minutes by adding galactose to the medium. Cells were then released from the G1 arrest by resuspending the culture in medium without alpha-factor and samples for FACS and whole-genome DNA sequencing were collected every 5 minutes up to 1 hour to monitor replication progression.

Overexpression of SSDDCS advances replication in the cell population, as measured by flow cytometry (Fig. 3.2A), which is consistent with previous observations from the lab[76]. In order to visualise replication dynamics genome-wide, samples for whole-genome DNA sequencing were collected at the same time-points (Fig. 3.1) and the ratio of mapped reads in the S-phase samples compared to the G1 sample was calculated for each genomic 1 kb bin following Batrakou et al[29] (see Methods for detailed protocol and analysis). T_{rep} was calculated by fitting a sigmoidal curve to the replication profile from each genomic bin and extracting the time (in minutes after G1 release) at which each bin is half-way from one copy to two copies (Fig. 3.2B, red dot). As described in the preface, these samples were collected by Dr. Mark Johnson, while I was responsible for the bioinformatic analysis of the data.

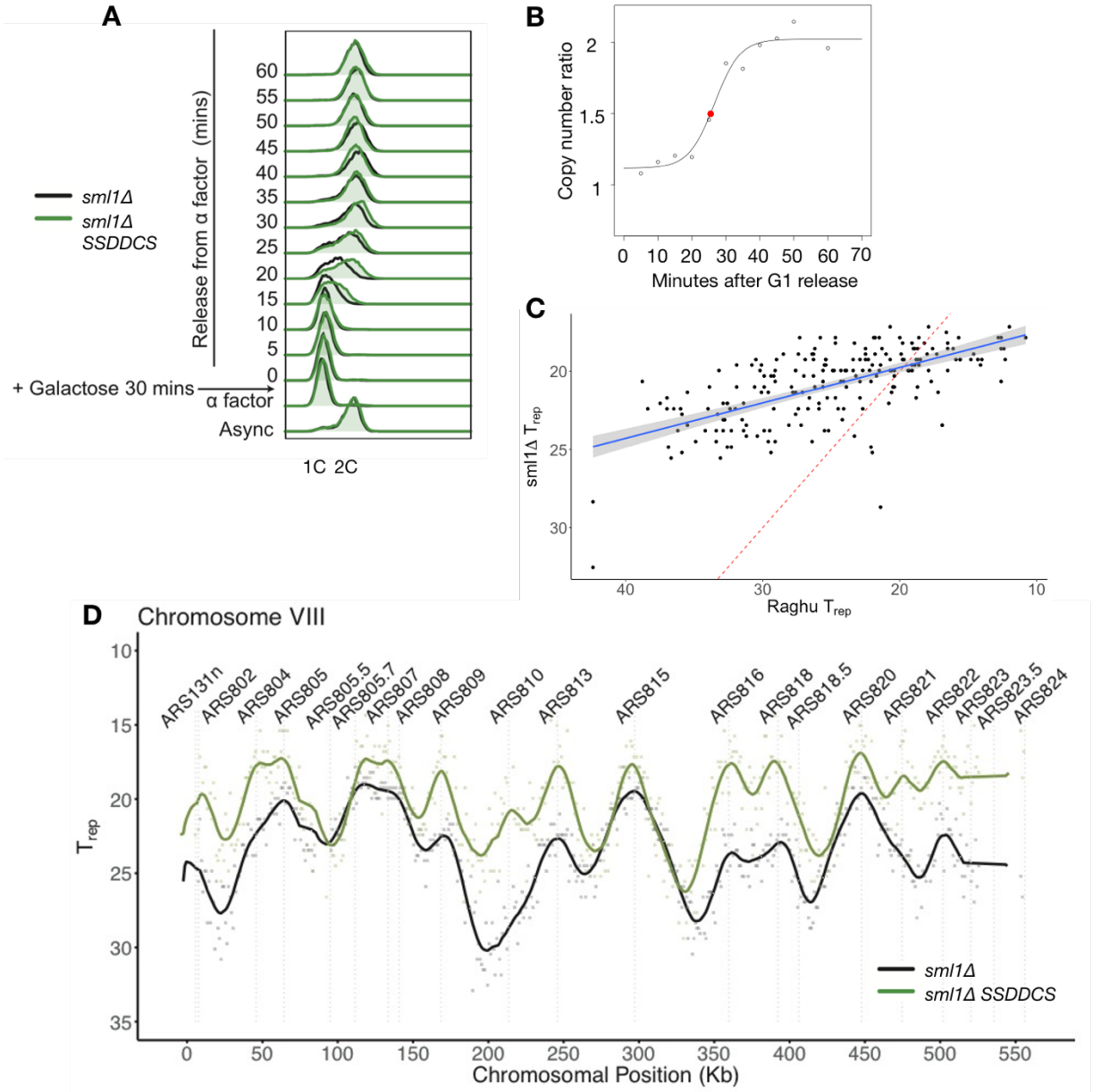


Figure 3.2 - Overexpression of six limiting factors advances replication timing. A – Bulk replication analysed by FACS for *sml1Δ* and *sml1Δ* SSDDCS strains. The SSDDCS strain starts S-phase faster compared to the control strain *sml1Δ* (notice the advance at 15-20 minutes post G1 release). 1C – one copy of the genome; 2C – two copies of the genome. **B** – Copy number ratio (ratio of mapped reads in S-phase samples to the non-replicating G1 samples) for one genomic 1 kb bin. T_{rep} was determined as the time of half-maximal replication (red dot, ~25 minutes for this particular bin). This value was calculated for every 1 kb bin in the genome. **C** - Scatterplot of origin T_{rep} values from *sml1Δ* strain vs T_{rep} determined by Raghuraman et al.[4] Despite the differences in absolute T_{rep} values, temporal order of replication is conserved. Blue line – linear regression, R^2 = 0.4069; red dashed line - equal T_{rep}, used to illustrate that, overall, origins fire earlier in *sml1Δ* compared to the strains used by Raghuraman et al. **D** - T_{rep} values were plotted along the corresponding chromosome positions to generate genome-wide replication profiles and smoothed using a moving average. Chromosome VIII is shown here as an example. The y-axis is flipped so that early replicating regions are at the top of the plot and late replicating regions at the bottom. The location of annotated origins (ARS) was overlayed to the profile and they align with peaks, as expected. Overall, all origins fire earlier in the SSDDCS strain.

To check that the control strain *sml1Δ* behaves as a wild-type, these origin T_{rep} values were compared with the values obtained by Raghuraman et al[4]. Comparison of *sml1Δ* data with the Raghuraman dataset shows that, despite the differences in absolute T_{rep} (on average origins fire 4 minutes earlier in *sml1Δ*), the temporal order of origin firing is maintained (Fig. 3.2C – blue linear regression line). One possibility for the differences in absolute T_{rep} are the different methodologies and growth conditions used: dense isotope transfer followed by DNA microarrays was used in Raghuraman et al. compared to whole-genome high-throughput sequencing used during this work. Moreover, in our experiments alpha-factor was used to synchronise the population, while in the Raghuraman study a temperature-sensitive *cdc7* mutant was used. Despite the different methodologies, the *sml1Δ* strain recapitulates the established temporal programme of origin firing (Fig. 3.2C). Replication profiles were generated by plotting the T_{rep} values along the corresponding chromosome position (Fig. 3.2D, raw data points) and smoothed using a moving average (Fig. 3.2D, lines). As expected from the FACS profiles, SSDDCS overexpression advances replication timing of whole chromosomes (Fig. 3.2D - profile of chromosome VIII).

In order to analyse the advance in RT genome-wide in an unbiased way, the T_{rep} values of all origins upon overexpression of SSDDCS were compared. As expected, the vast majority of origins fired earlier in the SSDDCS overexpression strain, while a small group fired at the same time or later (Fig. 3.3A). Moreover, origins fired 4 minutes earlier on average upon overexpression of the SSDDCS (Fig. 3.3B), which considering the short S-phase of budding yeast (15-20 minutes), accounts for an advance of approximately 20%. As described in the Introduction, RT is a highly regulated and robust process, and most studies attempting genome-wide perturbations reported a lower percentage of RT changes[164]. As such, the impact of SSDDCS overexpression can be considered to be highly significant.

In order to address whether early and late origins were equally affected by overexpression of limiting factors, origins were divided into quintiles according to

T_{rep}. This analysis demonstrated that late origins have a greater advance in RT compared to early origins upon SSDDCS overexpression (Fig. 3.3C-D).

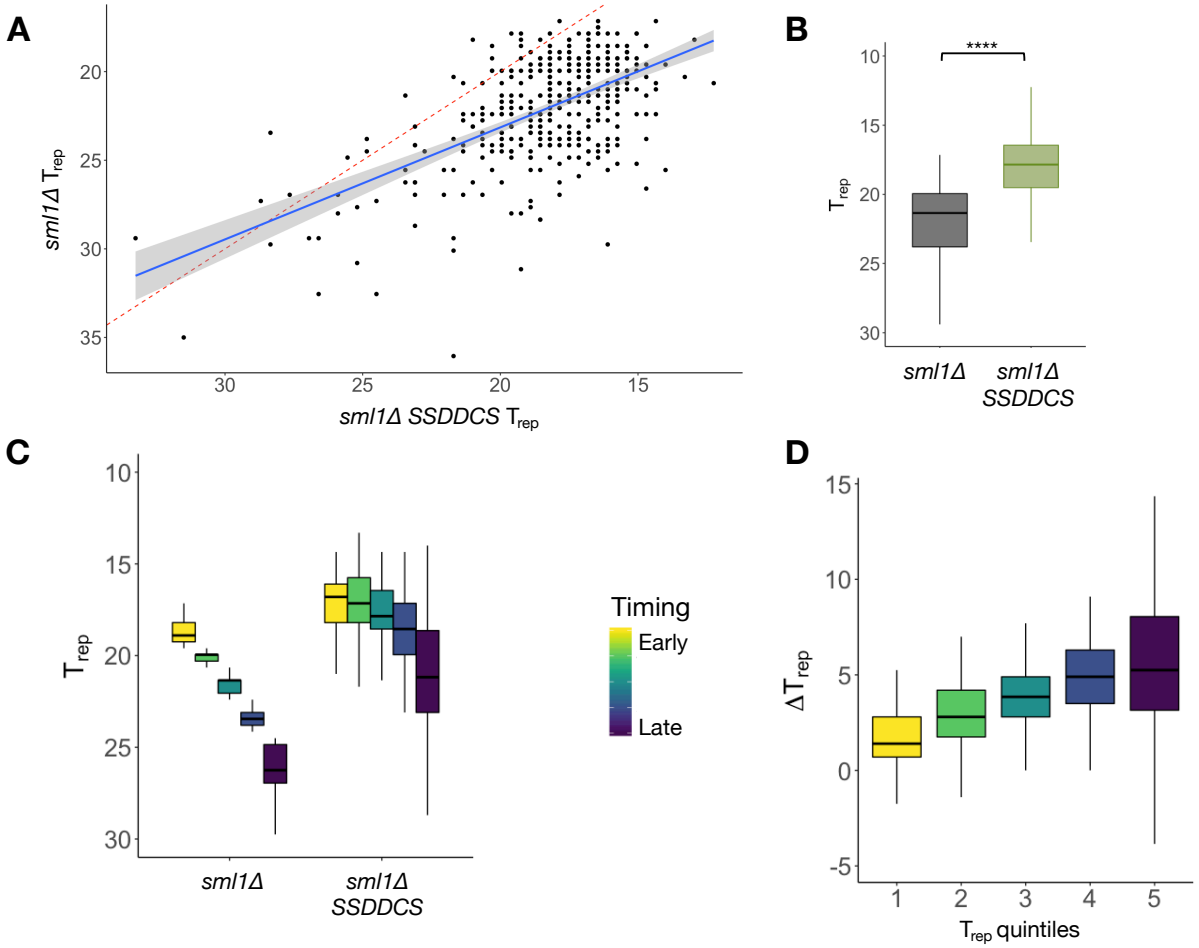


Figure 3.3 - Overexpression of six limiting factors advances RT genome-wide and late origins are more affected than early origins. A – Scatterplot of origin T_{rep} values for *sml1Δ* and SSDDCS strains. Almost all origins fired earlier in the SSDDCS strain. Blue line – linear regression, $R^2 = 0.3331$, used to illustrate the conservation of temporal order; red dashed line - equal T_{rep}, used to illustrate that, overall, origins fire earlier in SSDDCS (below red dashed line). **B** – Distribution of T_{rep} values for all origins in *sml1Δ* and *sml1Δ* SSDDCS. On average, origins fire 4 minutes earlier in SSDDCS. **** p-value < 2.2e-16, Welch Two Sample t-test. **C** - Same as B but origins were divided into quintiles according to T_{rep} values from *sml1Δ*. **D** - ΔT_{rep} (*sml1Δ* - *sml1Δ* SSDDCS) values for all origins divided into quintiles as in C. Later origins show greater ΔT_{rep} values, so their replication is more advanced compared to early ones.

Interestingly, despite the differences in absolute T_{rep}, the relative temporal order appears to be sustained in the SSDDCS strain (Fig. 3.3A – linear regression line and Fig. 3.3C). Early origins are still the earliest to fire, but the ΔT_{rep} of early origins is lower compared to late origins (Fig. 3.3D), possibly because there is a limit to even earlier activation of origins, possibly due to the requirement for S-phase CDK and DDK activation. Although late origins fire much earlier in S-phase (Fig. 3.3D), they are still later than early origins (Fig. 3.3C). This is also the case for early-replicating

centromeres, which are replicated with similar time in both strains (Fig. 3.4A-B), while telomeric regions which are late-replicating (regions within 50kb of chromosome ends) were significantly earlier replicated upon overexpression of the six limiting factors (Fig. 3.4C-D). These results show that RT can be dramatically advanced, genome-wide, but some RT differences are still observed between early and late origins. This may be because over-expression of these limiting factors is not penetrant enough to advance all origins equally, or other mechanisms that are independent of the concentration of initiation factors are also important to preserve RT, as described in the Introduction.

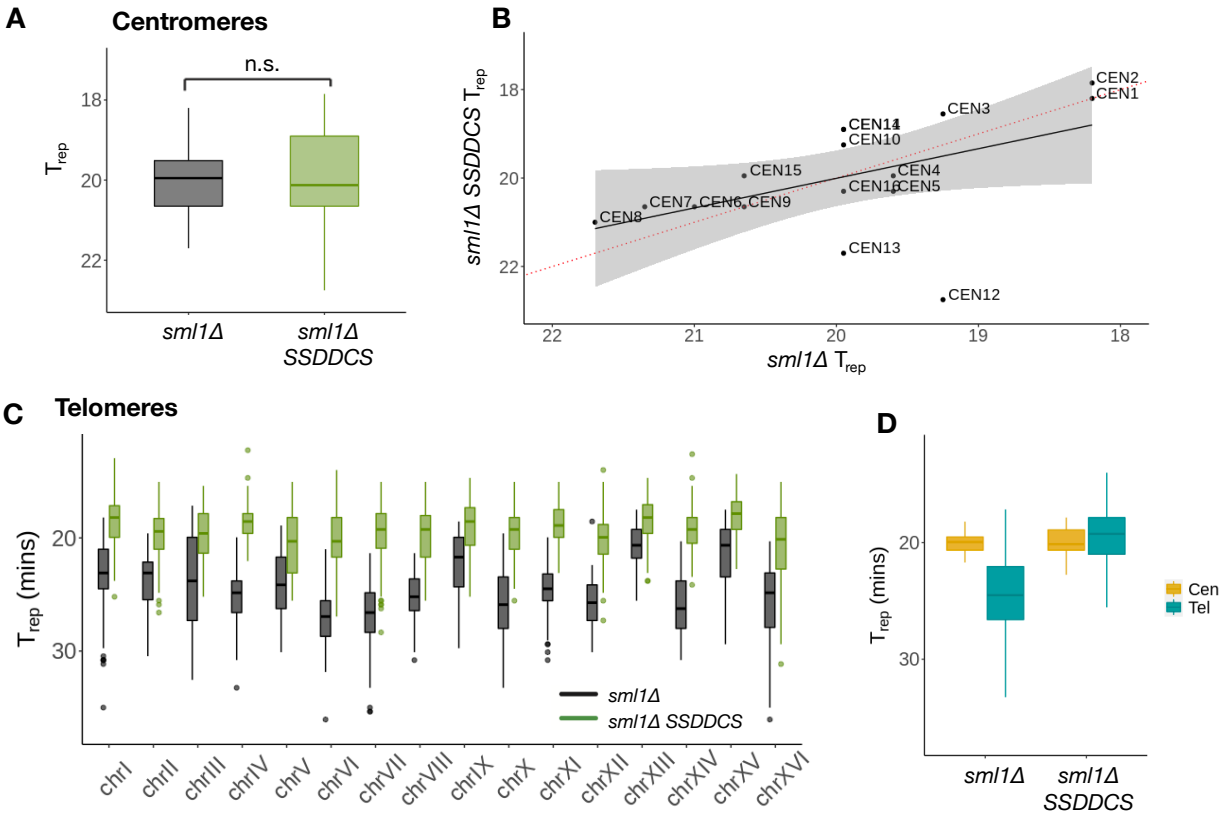


Figure 3.4 - Overexpression of six limiting factors advances RT of telomeres but does not affect centromeres, which remain early replicating. A – Distribution of T_{rep} values of the 16 centromeres. Centromeres remained early replicated and were not significantly affected by SSDDCS overexpression (n.s. – non significant, p-value = 0.9578, Welch Two Sample t-test). **B** – Scatterplot of centromeres T_{rep} values for each strain. Equal T_{rep} line (red dashed) overlaps with linear regression line (black, with confidence interval in grey). **C** - Distribution of T_{rep} values of telomeric regions (genomic bins within 50kb of chromosome ends), plotted by chromosome. All telomeric regions were significantly earlier replicated (**** p-value < 0.0001, Welch Two Sample t-test). **D** - Distribution of T_{rep} values of all telomeric regions and centromeres. In the control strain, centromeres were early replicated and telomeres late replicated, a fundamental feature of the RT programme. Upon SSDDCS overexpression, telomeres replicated earlier compared to the *sml1Δ* strain and at the same time as centromeres in the SSDDCS strain.

The SSDDCS conditional system provides a unique tool to investigate the impact of advancing RT in a single cell cycle, and its implications to the genome structure and function. Despite the several studies comparing RT and transcription, a complete understanding of the relationship between the two is still missing. As such, the SSDDCS system was used to investigate the impact of advancing RT on gene expression.

3.2 - Overexpression of limiting initiation factors affects gene expression during S-phase

In order to analyse the impact of an advanced RT on gene expression during S-phase, samples were collected for whole transcriptome sequencing (RNA-Seq) following the experimental set up from Figure 3.1. As described in the preface, these samples were collected by Dr. Mark Johnson, while I was responsible for the bioinformatic analysis of the data. This experiment was repeated 4 times, and Principal Component Analysis (PCA) was performed to address experimental covariates and batch effects (Fig. 3.5). The PCA plot shows that overall, biological replicates of the same time-point cluster together and apart from replicates from different time-points. Also, it illustrates the cell cycle regulated nature of gene expression, which seems to be a stronger clustering determinant compared to the differences between the two strains, i.e. the two strains cluster together in the same S-phase time-points (Fig. 3.5). This does not mean that there are no differences between the strains, but suggests that overall, cell cycle regulated genes are not affected and illustrates the periodic nature of cell cycle gene expression in budding yeast.

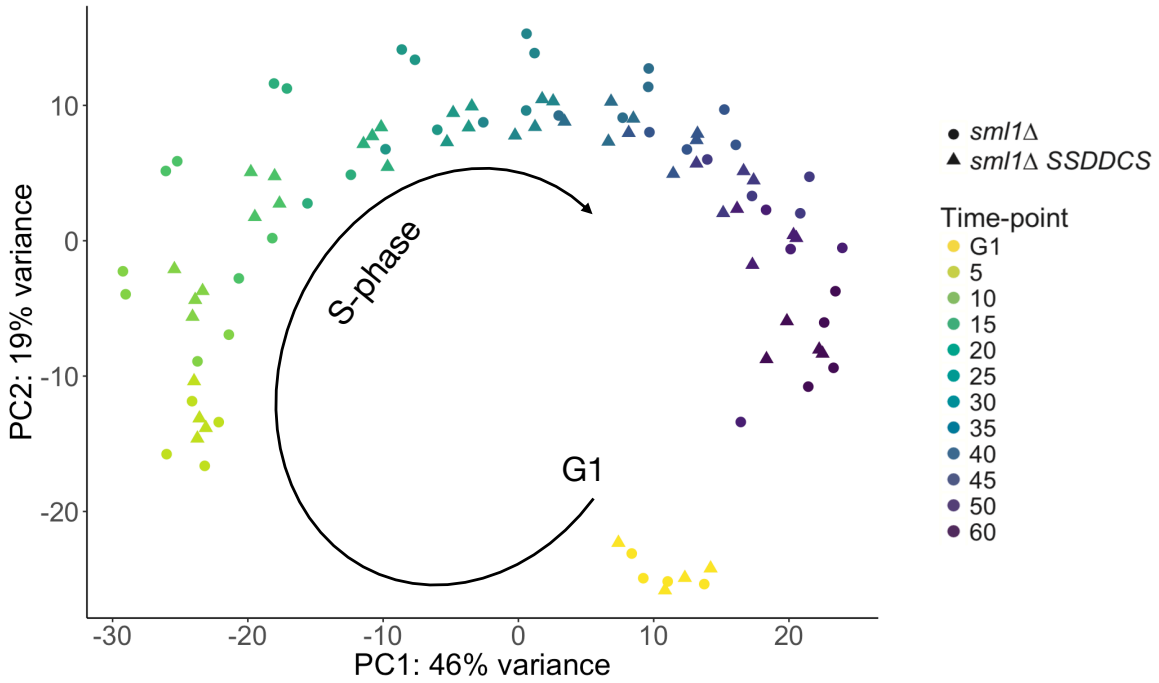


Figure 3.5 – PCA analysis of RNA-Seq samples. – PCA plot illustrating the clustering of individual replicates per time-point and strain. Samples are mostly clustered by time-point, illustrating the cell cycle nature of gene expression regulation in budding yeast and that overall, cell cycle regulated gene expression was not affected. N = 4

PCA analysis allows the identification of batch effects or experimental artifacts that can impact the results, and it is clear from the plot that this was not the case and that the results were reproducible between the 4 replicates (Fig. 3.5). Therefore, these datasets were used to explore the impact of an advanced RT on gene expression during S-phase.

To validate the RNA-Seq, the gene expression profiles of the six limiting factors were determined. As expected, the six limiting initiation factors were highly overexpressed at G1, and the overexpression was maintained throughout the time-course only in the SSDDCS overexpression strain (Fig. 3.6).

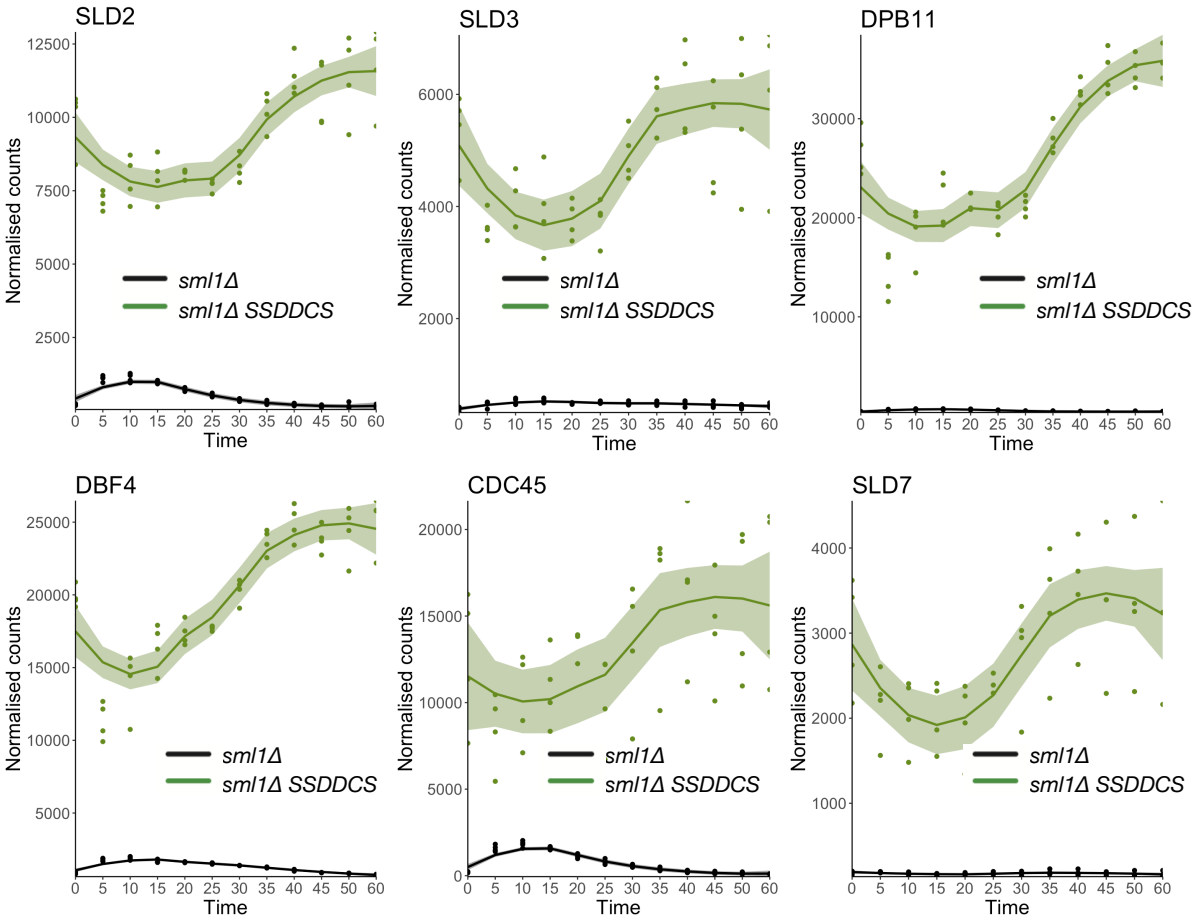


Figure 3.6 – SSDDCS are over-expressed at the mRNA level. – Normalised read counts per time-point for the six limiting factors Sld2, Sld3, Dpb11, Dbf4, Cdc45 and Sld7 (SSDDCS) using DESeq2 size factors. SSDDCS were highly overexpressed in G1 and remained over-expressed throughout and until the end of S-phase. The 4 replicates are shown for each time-point and strain together with the smoothing line generated using the loess method.

Upon confirmation of the SSDDCS overexpression at the mRNA level, these six genes were excluded from all downstream analyses. Then, differences in gene expression per time-point between the two strains were calculated using DESeq2[156]. Table 3.1 summarises this analysis, showing how many genes were significantly up or down-regulated (SSDDCS / *sml1⊿*) on each time-point, with an adjusted p-value or false discovery rate (FDR) < 0.01 (see Methods for detailed analysis). A very small number of genes was differentially expressed (DE) at G1 (16 and 26 up and down-regulated, respectively) but no particular biological process was associated with this group of genes, showing that the SSDDCS overexpression has minimal effects on the transcriptome during G1, ruling out downstream effects independent from changes in the RT programme. Interestingly, most changes took place during mid to

mid-late S-phase: there were not many genes affected at the early (5 to 15 minutes after G1 release) or late time-points (45 to 60 minutes after G1) (Table 3.1 and Figure 3.7). These results support the possibility that most changes in gene expression observed are a direct consequence of a dysregulated RT programme. Moreover, as cells progressed through S-phase, the changes in gene expression were progressively mitigated, such as by the end of the time-course there were almost no genes significantly affected (Table 3.1 and Fig. 3.7).

Time-point	Up-regulated	Down-regulated
G1	16	26
5	0	1
10	0	0
15	17	5
20	281	127
25	526	319
30	556	348
35	450	271
40	284	153
45	173	58
50	101	26
60	26	7

Table 3.1 – Differentially expressed genes per time-point. – Genes were considered up or down-regulated in the SSDDCS strain if the $\log_2$ normalised fold-change (SSDDCS / *sml1⊿*) was above or below 0, respectively and if the adjusted p-value or false discovery rate (FDR) was below 0.01 (DESeq2 Wald test). Cells are colour-coded by column based on the number of DE genes on each time-point.

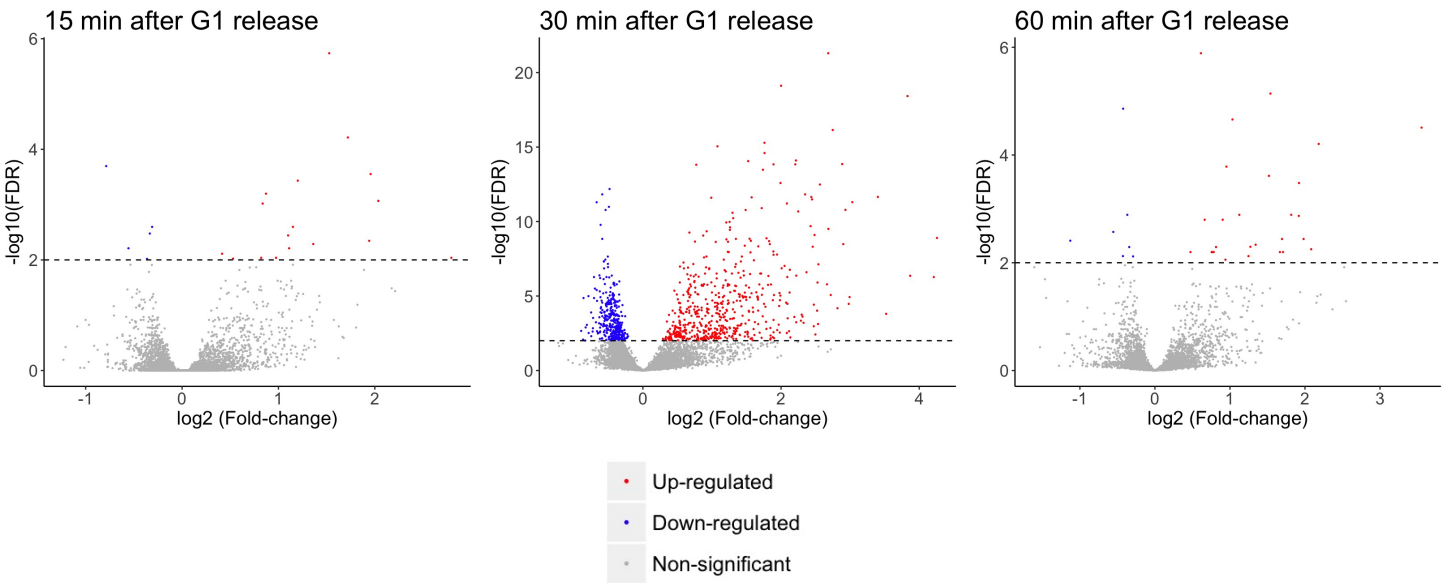


Figure 3.7 – Most changes in gene expression take place during mid to late S-phase. – Volcano plot for early (15), mid-late (30) and G2 (60) time-points. Fold-change differences in expression (SSDDCS / *sml1⊿*) and FDR were log normalised. Dashed horizontal line sets the minimum FDR required for a gene to be considered significantly differentially expressed (0.01). Up and down-regulated genes are coloured in red and blue respectively, and genes with no statistically significant changes are coloured in grey.

These observations support a direct but transient effect on gene expression, as gene expression patterns are re-established once DNA replication is finished.

The temporal resolution of this dataset facilitated an analysis of genes that show differential expression at more than one time-point, which could also increase the confidence in the expression change at that locus. In order to analyse expression patterns during S-phase, the DESeq2 likelihood ratio test (LRT) was used to identify genes which reacted differently between the two strains during the time-course: genes with a significant p-value from this test are those which at one or more time-points after G1 showed a strain-specific effect[156]. Then, k-means clustering was used to group genes in clusters with similar expression profiles (Fig. 3.8A-B).

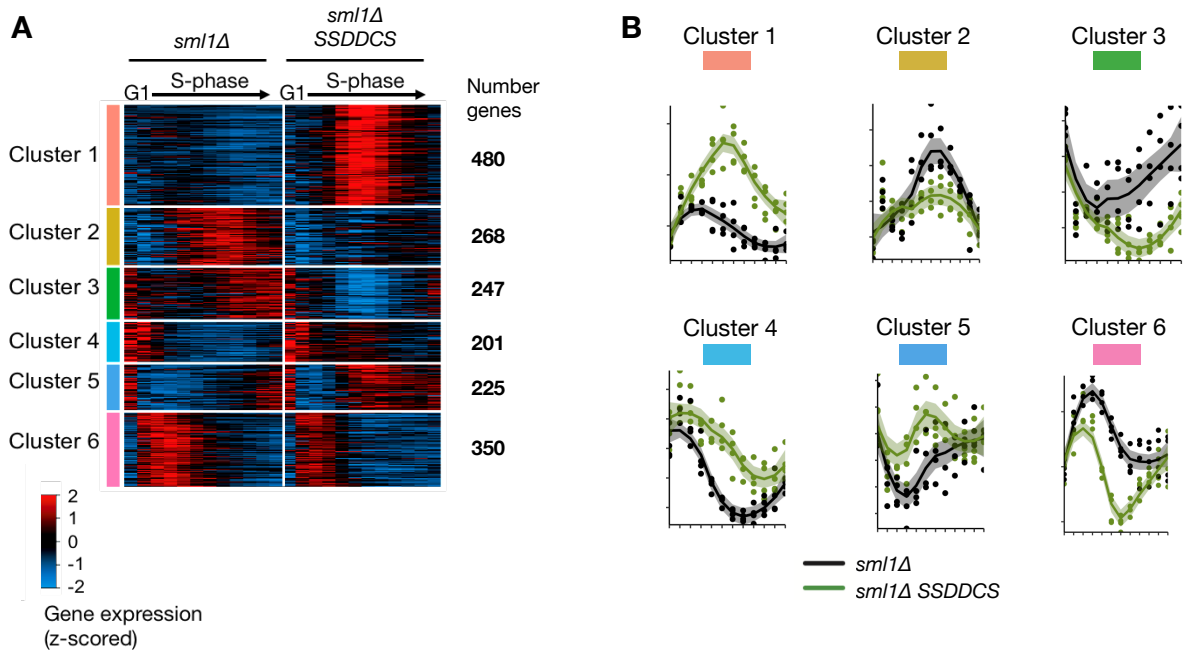


Figure 3.8 – Overexpression of limiting initiation factors has a heterogeneous effect on gene expression during S-phase. A - Heatmap of 1771 DE genes between the two strains in one or more time-points after G1 (FDR < 0.01 from DESeq2 LRT test). Each row corresponds to one individual gene and each column to one time-point (G1 to 60 from left to right, respectively). The two strains were separated using a white vertical line. Genes were clustered according to their expression profiles using k-means clustering. Each cluster is colour coded with a vertical bar on the left-side of the heatmap and separated from other clusters using horizontal white lines. The expression levels were z-scored and normalised by row, so the colours represent how far each value is from the mean of the values of the row. In sum, the time-points of maximal and minimum expression are coloured in red and blue, respectively. The total number of genes per cluster in indicated on the right side of the heatmap. **B** – Gene expression profiles of one example from each DE cluster (plotted as in Fig. 3.6), to illustrate the heterogeneous effect on gene expression. A colour bar representing each cluster was added to the top of the plots.

Using this approach, a total of 1771 genes were identified as differentially expressed (DE), representing ~27% of the genome. The k-means clustering allowed the identification of groups of genes with distinct patterns: cluster 1 genes are lowly expressed in the control strain and up-regulated in the SSDDCS strain, while cluster 2 genes have the opposite pattern. Cluster 3 genes are expressed in G1 and their expression drops during the early time-points but increases in later time-points, and these genes are down-regulated in SSDDCS. Cluster 4 is similar to cluster 3 in terms of expression dynamics, but these genes are up-regulated in SSDDCS. Cluster 5 genes are similar to 4, but their expression starts to increase in mid rather than late time-points. Finally, cluster 6 genes are lowly expressed in G1, followed by an increase during the early time-points and a drop in mid-late time-points, and

they tend to be up-regulated in the control strain. All groups have similar expression levels in the first and last time-point, except for cluster 4, whose expression drops in later time-points (Fig. 3.8).

The association of DE genes with particular genomic features such as distance to origins, telomeres and centromeres was also addressed. Compared to non-differentially expressed genes (NOTDE), cluster 1, 4 and 5 are located significantly closer to origins, while clusters 2, 3 and 6 are more distant (Fig. 3.9A). It is important to note that the median distance of non-differentially expressed genes to origins is the same as the median of the whole genome (8.9 kb), so DE clusters can be directly compared to the non-DE genes.

To determine if DE genes were significantly clustered in telomeric or centromeric regions, the proportion of genes within 50kb of telomeres or centromeres was compared for each cluster. Cluster 1 and 4 are the only groups with significantly more genes located in sub-telomeric regions compared to the genome average (Fig. 3.9B - 12% of all genes in the genome are in sub-telomeric regions compared to 23% and 18% of genes in cluster 1 and 4 respectively). Considering that genes in clusters 1 and 4 are up-regulated in SSDDCS (Fig. 3.8), it is possible that sub-telomeric heterochromatin silencing is affected in the SSDDCS strain. All clusters have the expected proportion of genes within 50kb of centromeres (Fig. 3.9C).

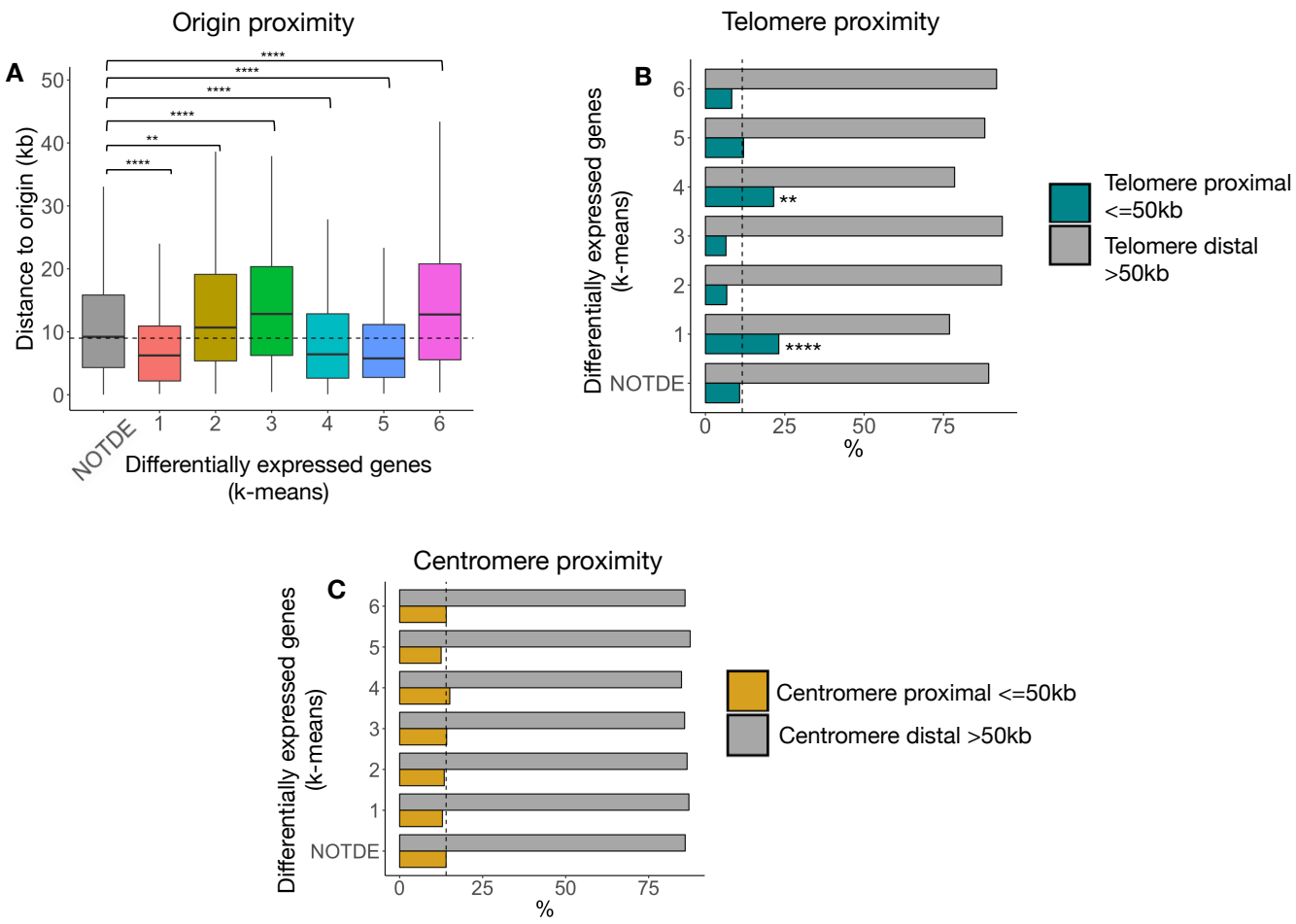


Figure 3.9 – Some DE clusters are associated with origins and telomeres, but none is associated with centromeres. A – Distribution of gene distances to the closest origin for every gene in the genome, split by k-means cluster. NOTDE represents all non-differentially expressed genes. Dashed horizontal line marks the genome-wide median gene distance to the closest origin = 8.9 kb. p-values are from pairwise comparisons of each k-means cluster versus the non-DE genes using Wilcoxon rank sum test. **B** – Proportion of genes which are within or without sub-telomeric regions (less or more than 50kb away from the closest telomere, respectively). Vertical dashed line marks the percentage of all genes located in sub-telomeric regions = 12%. p-values are from an exact binomial test. C - Proportion of genes which are within or without sub-centromeric regions (less or more than 50kb away from the centromere, respectively). Vertical dashed line marks the percentage of all genes located in sub-centromeric regions = 14%. All groups had the expected proportion of genes in sub-centromeric regions. **** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$.

Sub-telomeric regions were significantly affected by the advance in replication timing induced by SSDDCS overexpression (Fig. 3.4C-D), and two up-regulated clusters have an over-representation of genes in sub-telomeric regions (Fig. 3.9B), which suggests a potential time-related effect for some of the changes in gene expression observed in these clusters, which could involve the de-repression of sub-telomeric chromatin (see Discussion). On the other hand, all DE clusters have

the expected proportion of genes proximal to centromeric regions, which may reflect the fact that RT is not affected at centromeres (Fig. 3.4A-B).

As all the DE gene clusters have different proximities to origins compared to the whole-genome median distance (Fig. 3.9A), this might suggest a link between RT and gene expression for at least some loci. Indeed, all DE clusters follow a T_{rep} distribution consistent with their distance to origins, i.e. early replicating genes are located closer to origins and vice-versa (compare Fig. 3.9A and Fig. 3.10A). This is expected, as the timing of replication is directly associated with distance to origins. If the changes in gene expression were completely independent from RT, all or most DE clusters would have a similar T_{rep} distribution compared to the non-DE, as is the case for genes located within 50kb of centromeres, for example (Fig. 3.9C).

Considering that all DE clusters have (statistically) significant different distributions, these differences were explored because they could represent a biologically meaningful effect of RT on gene expression. It is important to stress however, that since this is an analysis of a set of genes this could include genes with similar RT to wild-type. Analysis of the T_{rep} distribution of DE clusters in the control strain, showed that cluster 1 genes are the earliest on average, while clusters 2, 3 and 6 are the latest (Fig. 3.10A).

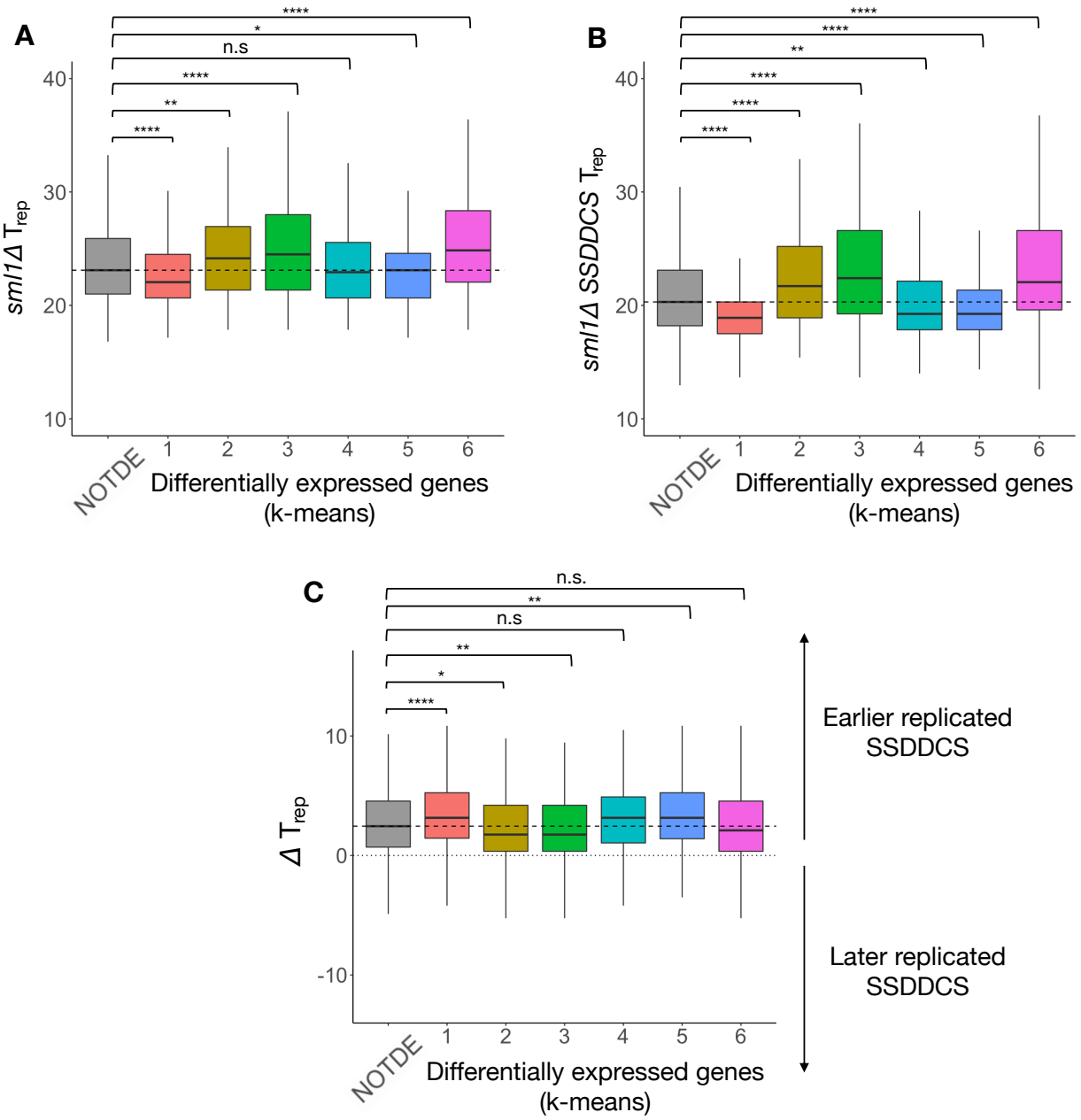


Figure 3.10 – DE clusters have different T_{rep} patterns. A – Distribution of T_{rep} values in the *sml1Δ* strain for all genes, divided by k-means cluster. NOTDE - all non-differentially expressed genes. Dashed horizontal line marks the median T_{rep} of all genes in the genome. p-values are from pairwise comparisons of each k-means cluster versus the non-DE genes using Wilcoxon rank sum test. **B** – Same as A but for T_{rep} values in SSDDCS. **C** – Distribution of ΔT_{rep} (*sml1Δ* - *sml1Δ* SSDDCS) for all genes, divided by k-means cluster. Dashed line marks the median ΔT_{rep} of all genes in the genome, while the dotted line marks 0 (no difference in T_{rep}). **** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, n.s. non-significant.

As expected, despite the overall earlier replication in SSDDCS strain, the temporal patterns of the DE gene clusters were maintained, whereby even in this compressed RT programme in this strain, cluster 1 genes are the earliest on average, while cluster 2, 3 and 6 are the latest (Fig. 3.10B)

In order to address the degree of change in RT for the different DE clusters, differences in T_{rep} between the two strains (ΔT_{rep} = *sml1Δ* - *sml1Δ* SSDDCS) were calculated. As expected, most genes across all clusters and non-DE were replicated earlier in the SSDDCS strain (Fig. 3.10C, dotted line). Moreover, non-DE genes do not have a significantly different ΔT_{rep} compared to the median effect on all genes in the genome (Fig. 3.10C – dashed line). However, some of the DE clusters were significantly affected in SSDDCS strain: cluster 1 genes were the most advanced genes, on average, followed by cluster 5. Clusters 2 and 3 were less affected compared to non-differentially expressed genes, while cluster 4 and 6 were not significantly different from non-DE genes (Fig. 3.10C).

In order to identify functional features common to the genes within each cluster, gene ontology (GO) enrichment analysis was performed using YeastMine[158] and the results visualised using REVIGO[165] (Fig. 3.11). This analysis allowed the identification of over-represented biological processes among different group of genes.

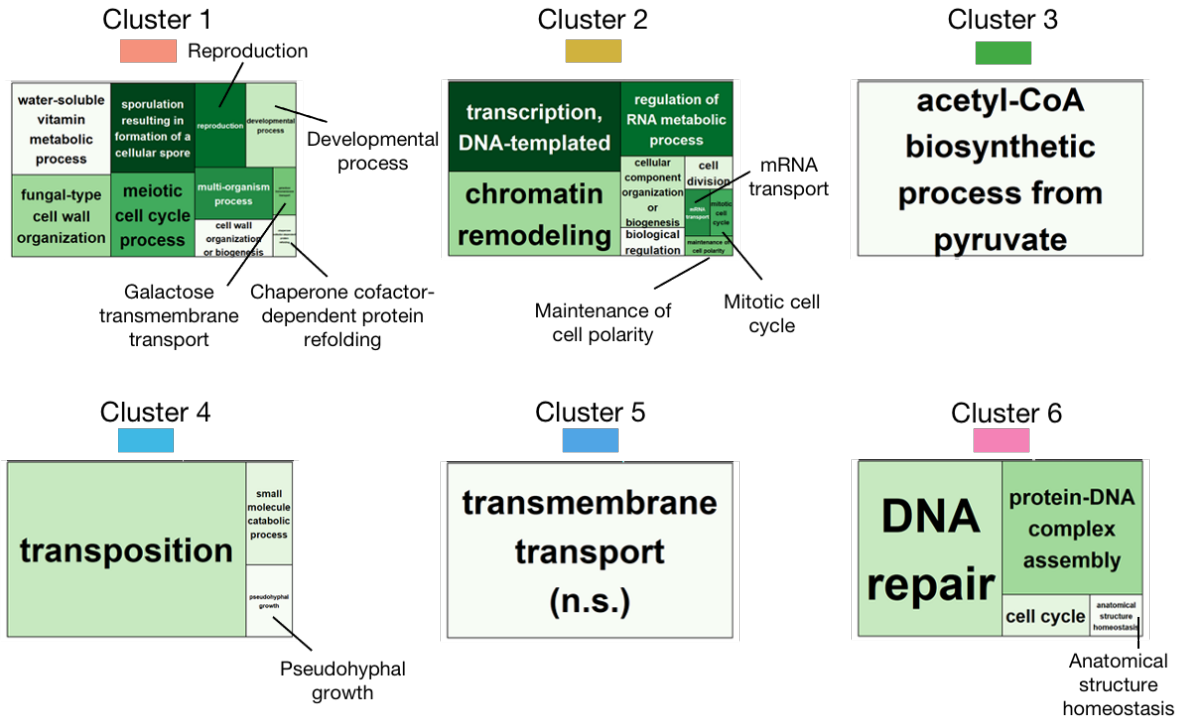


Figure 3.11 – Over-represented biological processes among differentially expressed clusters. Enrichment of gene ontology (GO) terms within each cluster of differentially expressed genes was determined using YeastMine[158], with a Holm-Bonferroni correction p-value cut-off of 0.05. Results were hierarchically clustered using REVIGO and visualised as a treemap. The size of each box is proportional to the aggregate p-value of all sub-categories corresponding to a parent GO term. There were no significantly enriched terms for cluster 5, so the first non-significant hit is shown.

Some of the enriched terms on cluster 1 represent families of sub-telomeric genes, such as the PAU gene family[166] (GO-term: fungal-type cell wall organization) and the THI5 gene family involved in thiamine biosynthesis[167] (GO-term: water-soluble vitamin metabolic process). Interestingly, it has been shown that thiamine gene expression is regulated by histone deacetylases such as Sir2, which silences heterochromatin at sub-telomeric regions[168]. This may suggest that some of these genes may be up-regulated in the SSDDCS strain due to defects in the silencing of heterochromatin (see Discussion). Cluster 2 and 6 had several biological processes over-represented, while cluster 5 had none (the first non-significant hit is shown). Cluster 3 had an over-representation for genes involved in acetyl-CoA biosynthesis, and cluster 4 for genes involved in transposition (Fig. 3.11).

A puzzling result was the over-representation of genes involved in meiosis and sporulation in cluster 1 (Fig. 3.11). Meiosis precedes spore formation and is triggered in budding yeast under poor nutrient conditions[132]. The budding yeast

lifecycle comprises an alternation between haploid and diploid stages: haploids of opposing types (a and alpha) mate to form diploids, while diploids form new haploids by sporulation. Both the haploid and diploid stage can divide by budding during the mitotic life cycle[169] (Fig. 3.12A). The diploid stage is the only stage that can undergo meiosis, but all the experiments presented in this thesis were done in haploids. As such, the meiotic gene expression programme should be turned off in these experiments, which is the case in the control strain (Fig. 3.8). However, important activators of the meiotic gene expression programme are part of cluster 1 and as such are up-regulated in the SSDDCS strain, such as *NDT80* and *IME2*[170,171] (Fig. 3.12B).

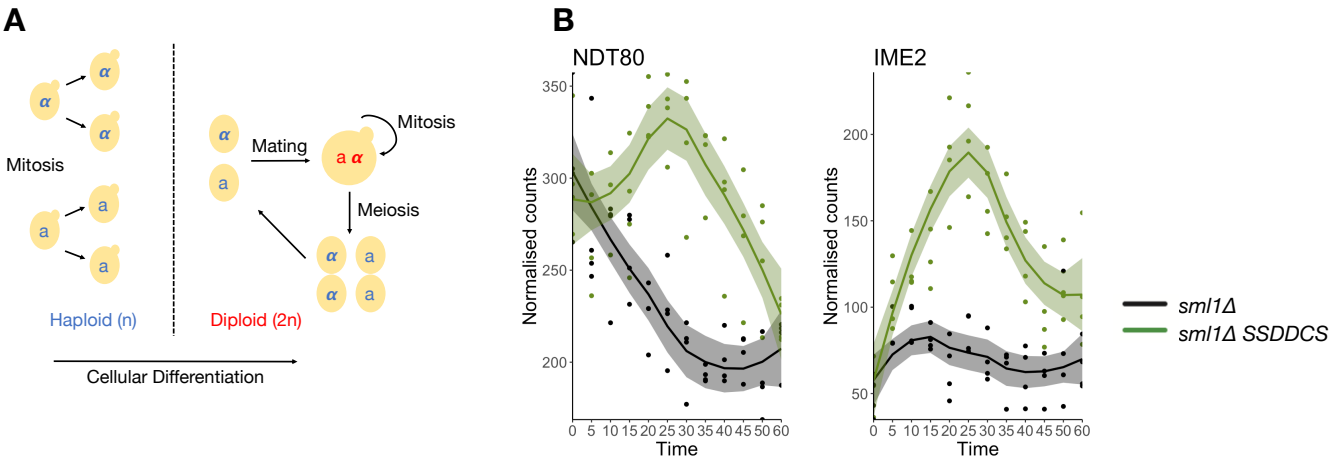


Figure 3.12 – Meiotic genes are expressed upon SSDDCS overexpression. A – Budding yeast life cycle. Haploids of opposite types (a and α) mate to form diploids, and diploids form haploids by sporulation. Both haploids and diploids divide by budding during the mitotic life cycle. The diploid stage is the only one with two copies of the genome and as such, the only type that can undergo meiosis. **B** – Key regulators of meiosis are overexpressed in the SSDDCS strain.

Considering that cluster 1 genes are: 1) up-regulated in SSDDCS (Fig. 3.8), 2) significantly closer to origins (Fig. 3.9A), 3) on average more advanced compared to any other DE cluster (Fig. 3.10C) and 4) enriched for genes involved in meiosis and sporulation (Fig. 3.11), which shouldn't be expressed in haploids, these genes were used to test whether any of the observed changes in gene expression was a direct consequence of an advance in RT.

3.3 - Direct effect of replication timing on gene expression

Overexpression of SSDDCS causes a genome-wide advance in RT, and affects the gene expression of many genes in a single cell cycle, but it is not clear from this whether there is a direct interplay between gene expression and RT. In order to determine causality, the SSDDCS overexpression system was combined with origin deletion experiments. This is possible in budding yeast because origins are defined by specific sequences (described in the Introduction) allowing local modulation of RT. By deleting individual origins in the SSDDCS strain, RT of specific loci can be delayed while the rest of the genome remains advanced. Then, samples can be collected for reverse transcription quantitative PCR (RT-qPCR) to analyse gene expression of individual genes. *IME2* and *NDT80* loci were selected because these genes represent key regulators of meiotic gene expression, have a significantly advanced T_{rep} (3.15 and 5.6 minutes, respectively) and are located relatively close to origins (closest origin is 6.1 kb and 2.1 kb away from *IME2* and *NDT80*, respectively).

Budding yeast strains overexpressing SSDDCS in combination with removal of the two closest origins to *IME2* (ARS1008 and ARS1009 - Fig. 3.13A) or *NDT80* (ARS816 and ARS818 - Fig. 3.13B) were used to determine if a local delay on RT re-establishes the expression patterns of *IME2* and *NDT80*. The two closest origins to each gene were removed to make sure that RT of these loci was significantly delayed, followed by collection of samples for whole-genome DNA sequencing and RT-qPCR as described previously. As described in the preface, these samples were collected by Dr. Mark Johnson, while I was responsible for the bioinformatic analysis of the data. As expected, removal of these origins significantly delayed RT in these locations while the rest of the genome remained advanced (Fig. 3.13 – C).

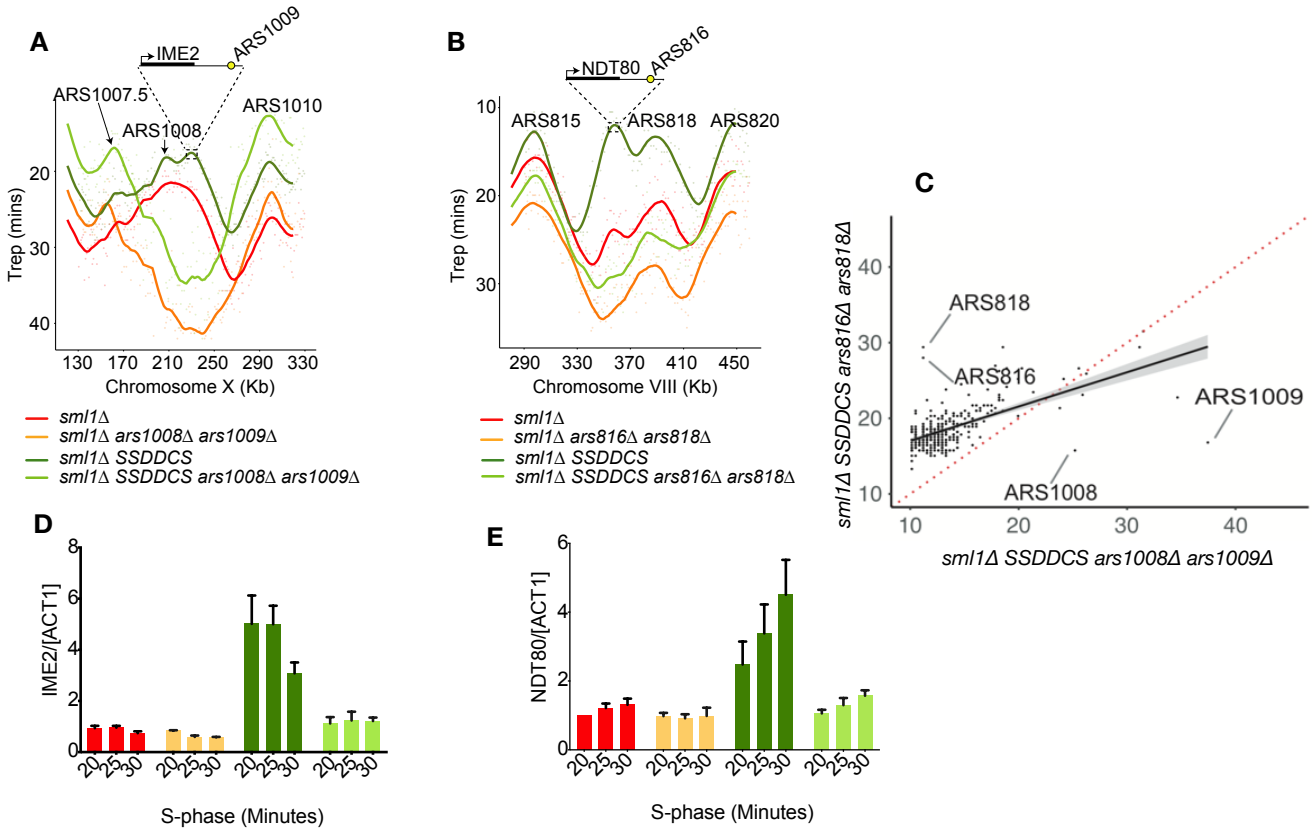


Figure 3.13 – Replication timing affects gene expression of *IME2* and *NDT80* directly. A – Replication profiles of *IME2* locus in *sml1Δ* and *sml1Δ* SSDDCS with and without the two closest origins to *IME2* (ARS1008 and ARS1008). Removal of these origins significantly delays replication timing of this location. **B** – Same as A, for *NDT80* locus (origins removed, ARS816 and ARS818). **C** – Scatterplot of T_{rep} values for all origins in the two SSDDCS overexpression strains with deleted origins (the two light green curves from A and B). All origins have comparable T_{rep} values, except for the genomic regions comprising ARS1008/ARS1009 and ARS816/ARS818 which were delayed on the strain where they were removed (*sml1Δ* SSDDCS ars1008Δ ars1009Δ and *sml1Δ* SSDDCS ars816Δ ars818Δ, respectively). Black line – linear regression, red dashed line - equal T_{rep}. **D** – Expression of *IME2* at time-points 20, 25 and 30 after G1 release on strains from A, measured by RT-qPCR. **E** – Same as D for NDT80 on strains from B. Gene expression is normalised to actin (ACT1). N = 3.

Comparison of the two SSDDCS strains with deleted origins in either location showed that these were the only affected regions, while the remaining origins have comparable T_{rep} values between the 2 strains (Fig. 3.13C, *IME2* origins – ARS1008 and ARS1009; *NDT80* origins – ARS816 and ARS818). Replication profiles were generated as previously described, which confirmed the significant delay in the loci where origins were deleted (Fig. 3.13A and B, compare light green and dark green). As a control, these origins were also deleted in a *sml1Δ* background, and these

regions were also delayed compared to *sml1Δ,* as expected (Fig. 3.13A and B, compare orange and red).

The expression of *IME2* and *NDT80* in the origin deletion strains was measured using RT-qPCR at time-points 20, 25 and 30, which represent the time-points with highest fold-change differences detected by RNA-Seq (Fig. 3.12B). As expected, the qPCR data agrees with the RNA-Seq, as these genes were overexpressed in the SSDDCS strains at the designated time-points (Fig. 3.13D and E, dark green bars). Strikingly, the delay in replication of these loci, caused by the removal of individual origins, restores the normal expression levels (Fig. 3.13D and E, light green bars). These results show that the activation of these meiotic genes is likely to be a direct consequence of their earlier replication, and are not due to downstream effects of the genome-wide advance in RT or over-expression of 6 replication factors. The expression of these genes was not affected in *sml1Δ* strains with or without the proximal origins, as expected (Fig. 3.13D and E, orange and red bars).

As described in the Introduction, the expression of some genes is directly affected by an increase in copy number during replication, such as the histone genes[102]. Voichek et al. found that on average, expression of the earliest S-phase genes is buffered against copy number changes through Rtt109 (a histone acetyltransferase), in order to maintain expression homeostasis during S-phase (described in Introduction)[106]. To rule out an effect of copy number on *IME2* and *NDT80* expression, the *rtt109Δ* RNA-Seq data from Voichek et al. was analysed. Notably the Voichek et al. dataset is similar to the one used in this thesis: samples for gene expression analysis by RNA-Seq were collected during a synchronous S-phase every two minutes after alpha-factor release up to 60 minutes.

However, *IME2* was below the detection threshold for every time-point, while *NDT80* was detectable in only five time-points in the wild-type and two time-points in the *rtt109Δ* mutant (Fig. 3.14A). These results are expected and agree with the RNA-Seq data generated in this thesis, as these genes are not normally expressed in haploid cells during mitotic S-phase. Moreover, an attempt to rule out copy number effects for all genes from cluster 1 was unsuccessful for the same reason (Fig. 3.14B). Similar results were observed for the remaining clusters (not shown).

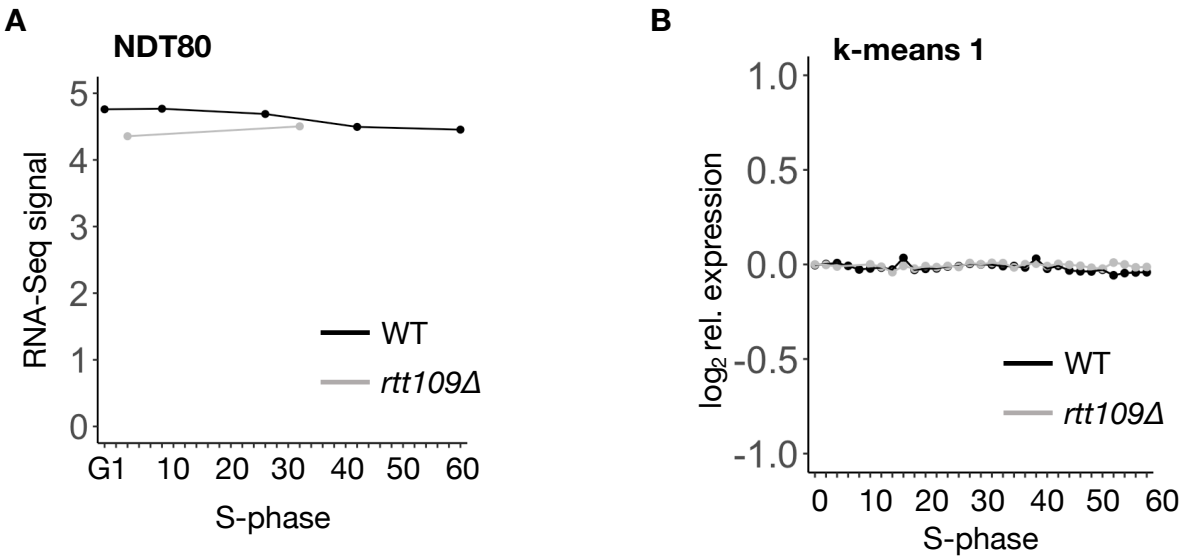


Figure 3.14 – A copy number effect cannot be ruled out using *rtt109Δ* data from Voichek et al. A – *NDT80* expression during S-phase in wild-type and *rtt109Δ* cells from Voichek et al.[106]. In most time-points, *NDT80* expression was below the detection threshold and expression levels were low in the detected time-points. The RNA-Seq signal is log transformed and represents the number of reads mapped to this locus (data was normalised so that the total signal was the same in each sample). **B** – Average relative expression of cluster 1 genes (log_2 S-phase expression / G1 expression) in the same dataset. As expected, these genes are lowly expressed in haploid cells during S-phase, making it difficult to rule out a copy number effect using this data.

The fact that the average expression of cluster 1 genes and *NDT80* is not affected in *rtt109Δ* mutant could potentially suggest that these genes are not buffered by this mechanism, but the simpler explanation is that these genes do not need to be buffered because they are not expressed in haploid cells during mitotic S-phase.

Still, a copy number effect on *IME2* and *NDT80* expression due to SSDDCS overexpression can most likely be ruled out for the following reason: the differences in expression between *sml1Δ* and *sml1Δ SSDDCS* were significantly greater than 2-fold for both genes (Fig. 3.13D and E, dark green bars), which wouldn't be the case if the up-regulation of these genes was completely dependent on an increase in copy number (i.e., these genes would only be 2-fold up-regulated). As such, these results suggest a timing effect on gene expression of these genes which is mostly independent from an increase in copy number.

Overall, the results presented during this chapter show that the genome-wide advance in RT caused by overexpression of limiting factors has an impact on the expression of ~27% of all budding yeast genes in a single cell cycle (Fig. 3.8A). Moreover, different DE clusters with distinct expression profiles were identified, illustrating the heterogeneity of the impact of RT on gene expression and the complex relationship between the two (Fig. 3.8B). Moreover, some DE clusters were associated with genomic features such as origins of replication and telomeres (Fig. 3.9A-B), had different degrees of advance in T_{rep} (Fig. 3.10) and important mediators of sporulation, a cellular differentiation event in budding yeast, were activated when cells were not ready to undergo this transition (Fig. 3.12). Finally, origin deletion experiments showed that for some genes, the impact of RT on gene expression is likely to be direct, as the local delay of RT in these loci was enough to restore wild-type expression levels (Fig. 3.13). Therefore the SSDDCS overexpression system in combination with the temporal resolution of our experimental design provides an unique opportunity to address the impact of an advanced RT on gene expression and the establishment of the chromatin landscape in S-phase.

Chapter 4 – Effect of a perturbed replication timing programme on the chromatin landscape

4.1 - MNase-Seq as a tool to analyse the nucleosome landscape genome-wide with base-pair resolution

Considering the close link between RT, gene expression and chromatin, the SSDDCS overexpression system was used in combination with MNase-Seq to address the impact of the global RT advance on the chromatin landscape. MNase-Seq stands for chromatin digestion with micrococcal nuclease (MNase) followed by high throughput sequencing. MNase is an enzyme that digests naked DNA (i.e. DNA that is not bound by proteins), allowing the isolation of DNA fragments which are protected by proteins, such as histones (Fig. 4.1A), and the generation of genome-wide nucleosome profiles with base-pair resolution.

To generate nucleosome profiles, the MNase-Seq protocol from Nocetti and Whitehouse (2016), which was used to analyse nucleosome movement during the yeast metabolic cycle[155], was optimised. Different fragments can be isolated depending on the MNase digest conditions, so this step was adjusted so that roughly 80% of the isolated fragments were 150bp in length, which corresponds to the size of single nucleosomes (i.e. length of DNA wrapped around the histone octamer - Fig. 4.1B). The next step on the protocol involves library preparation for sequencing, which includes the ligation of adapter sequences to the fragments ends, increasing the size of the isolated DNA fragments to approximately 270bp, (150bp + two 60bp adapters - Fig. 4.1C). Upon sequencing, adapter sequences were computationally removed and the single nucleosome fragments were sequenced and mapped to the reference genome (Fig. 4.1D, insert size – size of sequenced DNA fragment between the adapters). Peak calling was then performed in the mapped reads to identify nucleosome locations. DANPOS was used to identify MNase-Seq peaks and compare nucleosome position and occupancy between the two strains[159]. A detailed protocol and analysis can be found in the Methods section.

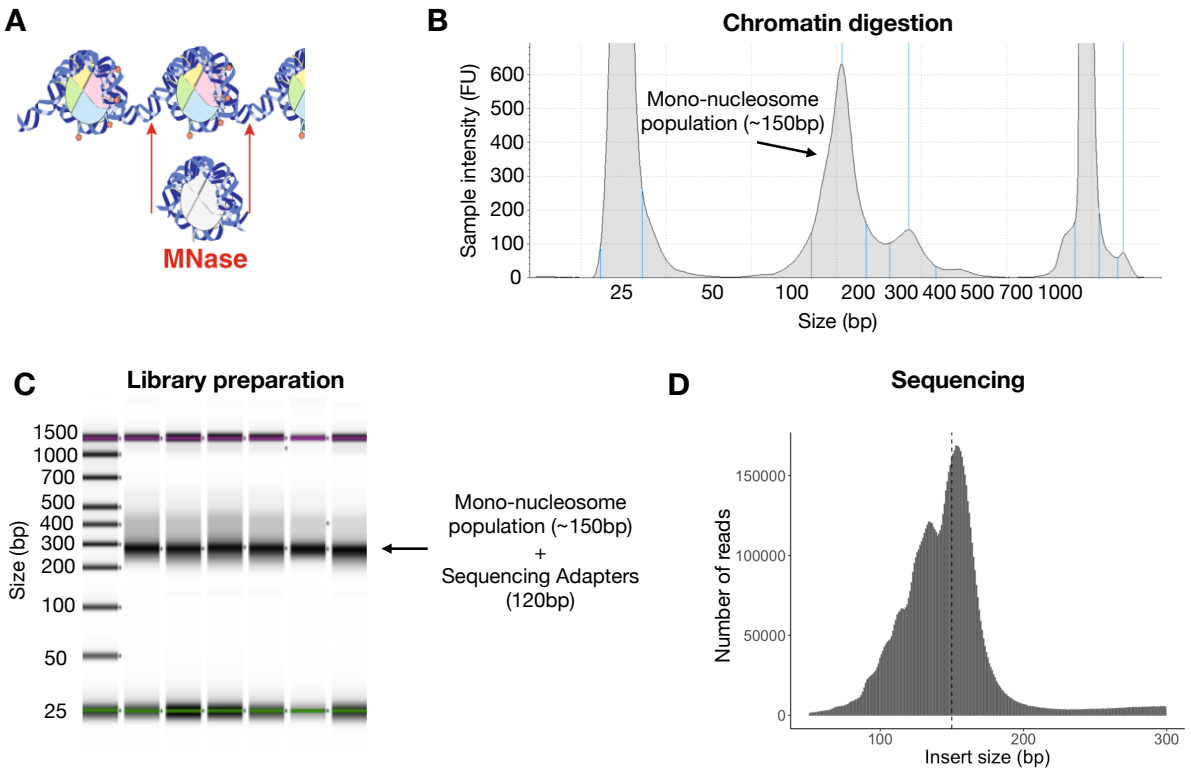


Figure 4.1 – MNase-Seq allows the isolation of DNA fragments bound by DNA binding proteins, such as histones, corresponding to individual nucleosomes. A – Micrococcal nuclease (MNase) digests naked DNA, which is not bound by proteins, such as the linker DNA connecting individual nucleosomes. After a proteinase treatment, these fragments are isolated and sequenced to generate whole-genome nucleosome profiles. **B** - Agilent D1000 ScreenTape quantification from one DNA sample after MNase digestion optimised to isolate mono-nucleosome fragments (~150bp). Notice the small peak of di-nucleosomes (~300bp), which is expected, as an attempt to completely digest this population could lead to the digestion of the mono-nucleosome population. The abundant populations at 25bp and >1000bp correspond to the lower and upper markers respectively, used to determine the size of the isolated fragments. Chromatin was digested with 90U of MNase for 3 minutes at 37°C. This distribution was similar across all samples. **C** - Agilent D1000 ScreenTape gel from six DNA samples after MNase digestion and library preparation. After adapter ligation, the average fragment size is approximately 270 bp, corresponding to mono-nucleosomes with adapters ligated on both ends (~150 + 60 + 60 = ~270 bp). This distribution was similar across all samples. **D** – Distribution of number of reads per insert size for one of the samples. As expected, the distribution peaks at approximately 150bp (dashed vertical line), further confirming that single nucleosome sized fragments were enriched and sequenced. This distribution was similar across all samples.

To study the impact of advancing RT on the nucleosome landscape, MNase-Seq was used in combination with the SSDDCS overexpression system and time-course resolution to generate nucleosome profiles during S-phase (Fig. 4.2A). To validate our MNase-Seq data I plotted the average nucleosome profiles of all genes (centred

on the transcription start site (TSS)) and of all origins (centred at the ARS consensus sequence (ACS)) to determine how well the G1 dataset recapitulated the well-described nucleosomal patterns of these sites (Fig. 4.2B and C). Consistent with other datasets, our data shows that promoters have a well-positioned nucleosome downstream of the TSS (+1 nucleosome), and a nucleosome depleted region upstream of the TSS[172], which is important for the binding of proteins required for initiation of transcription, such as transcription-factors and the RNA polymerase machinery (Fig. 4.2B), while origins have an asymmetric nucleosome free region flanked by well positioned nucleosomes, as described[48] (Fig. 4.2C). Figure 4.2B/C provides confidence that our MNase-Seq data is of sufficient quality to recapitulate established nucleosome profiles *in vivo*.

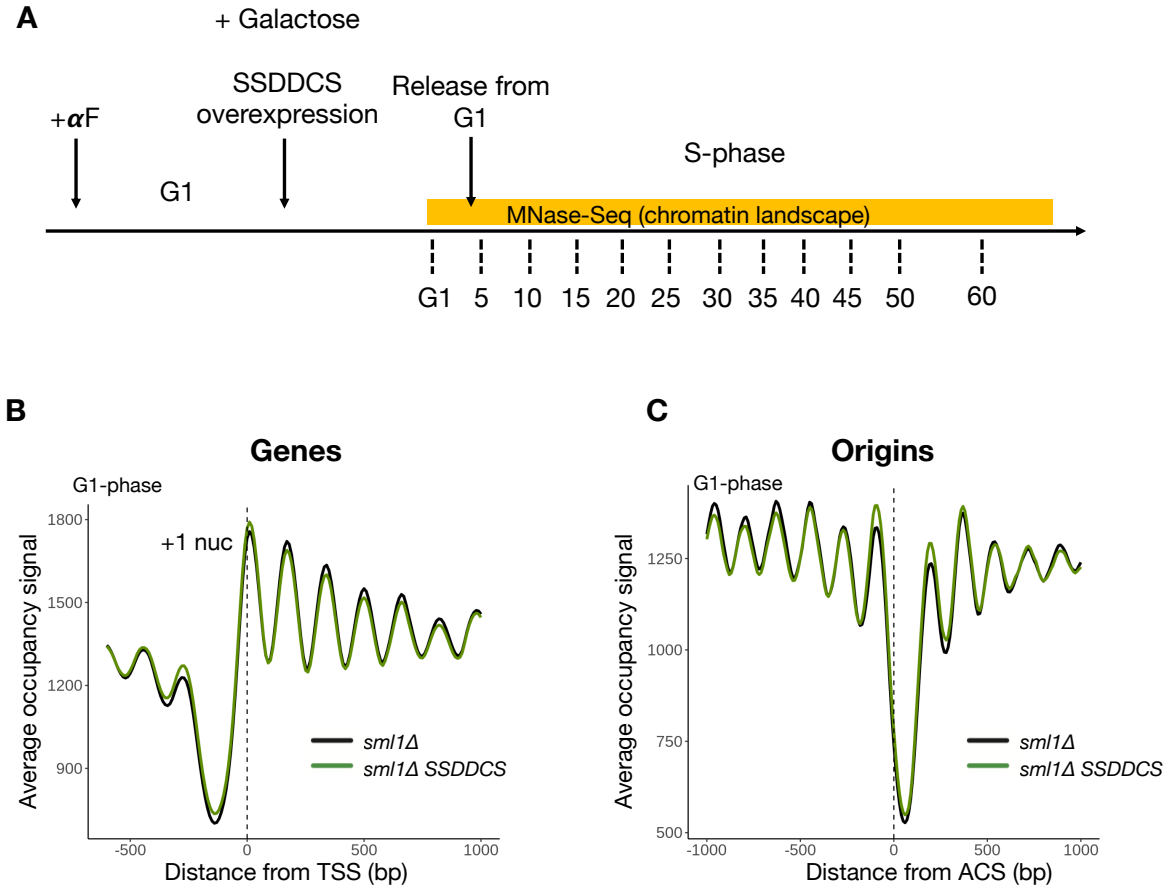


Figure 4.2 – MNase-Seq allows the study of the nucleosome landscape genome-wide. **A** – Experimental system described in previous sections, where a population overexpressing the limiting initiation factors is released into a synchronous S-phase. Samples were collected every 5 minutes for MNase-Seq. **B** – Average nucleosome profile for all genes at G1 time-point. A window of 500bp upstream and 1000bp downstream of the TSS (dashed vertical line) was used. **C** – Same as B, but for origins of replication. A window of 1000bp upstream and downstream of the ARS consensus sequence or ACS (dashed vertical line) was used. At the G1 time-point, the profiles from the two strains are nearly identical, as expected. Data is from 3 independent biological replicates.

The G1 average profiles of genes and origins in the *sml1Δ* strain were identical to the SSDDCS strain (Fig. 4.2 B-C), suggesting that SSDDCS doesn't cause major effects on chromatin outside S-phase. As such, this dataset was used to explore time-dependent effects on the chromatin landscape during S-phase.

The number of nucleosomes identified in each time-point varied between 68000 and 70000 (Fig 4.3A), which is in agreement with previous MNase-Seq studies[172]. There was a small variation on the total number of nucleosomes during S-phase, which dropped during early S-phase and increased during later time-points (Fig. 4.3B). This could be explained by the fact that nucleosomes have to be incorporated into the newly synthesised DNA during S-phase, and actively replicated regions will experience a brief nucleosome depletion. Nucleosome numbers are then re-established by the end of S-phase, as replication finishes and chromatin matures (Fig. 4.3B – the number of nucleosomes at G1 and 60 min is the same in the *sml1Δ* strain).

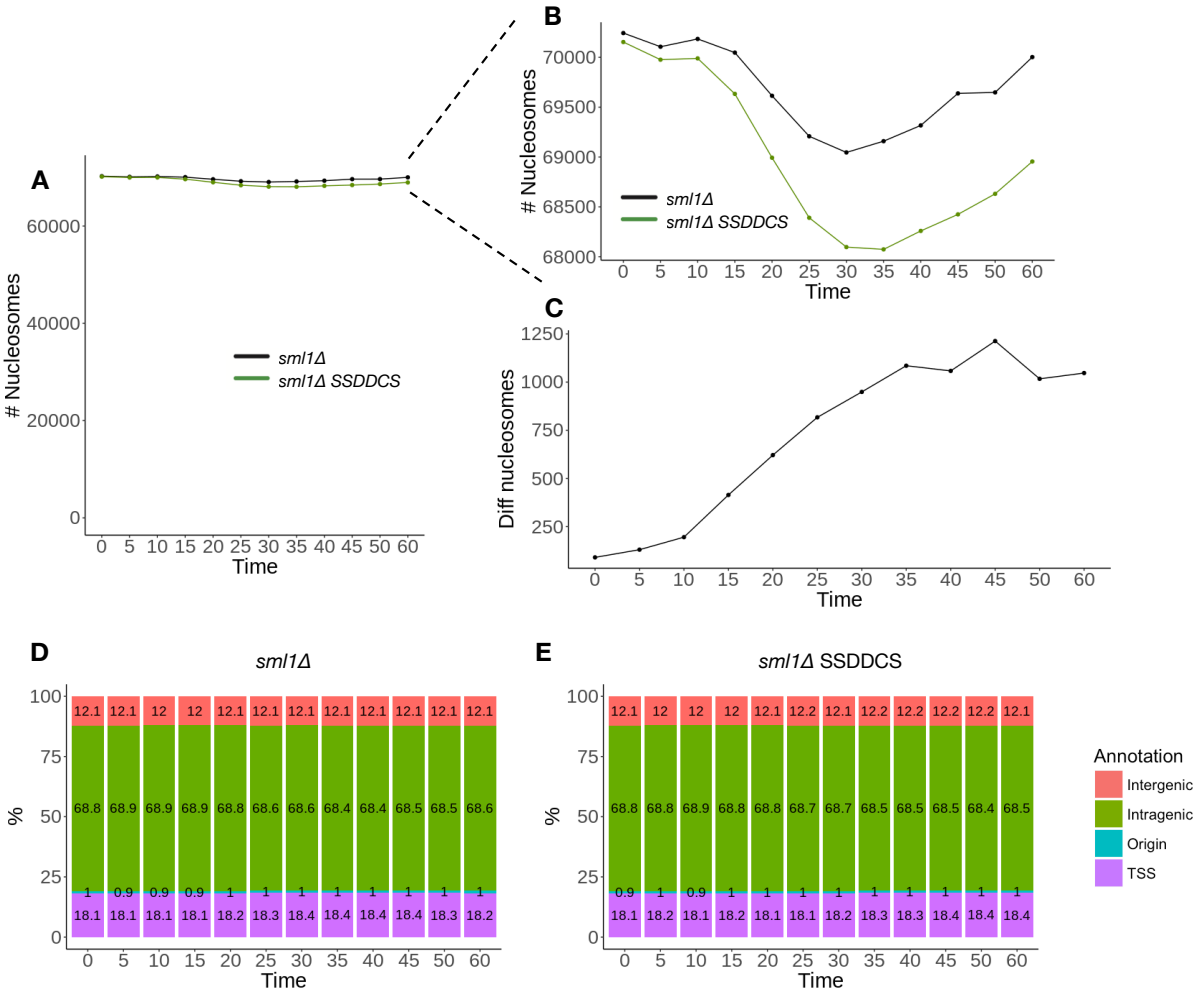


Figure 4.3 – The total number and distribution of nucleosomes across the genome is not affected upon SSDDCS overexpression. A – Total number of nucleosomes (MNase-Seq peaks) identified for each strain during the time-course. **B** – Zoom in on the plot from A to illustrate that the total number of nucleosomes varies slightly during S-phase, with a small decrease in early time-points and increase in later time-points. **C** – Absolute difference in total number of nucleosomes (*sml1Δ* – SSDDCS) per time-point. **D** – Proportion of promoter (TSS), intergenic, intragenic and origin nucleosomes in the *sml1Δ* strain. **E** – Same as D in the SSDDCS strain. Numbers within the bars represent the percentage of each category.

A small but consistent difference was present between the two strains in mid and late time-points (Fig. 4.3B-C), despite the equivalent number of nucleosomes in G1. Moreover, it seemed that the SSDDCS strain was not able to re-establish the total number of nucleosomes at the end of S-phase (Fig. 4.3B). The fact that there were no differences between the strains in G1 and early time-points could suggest a potential effect of the advanced replication on the chromatin landscape. Higher rates of initiation during early S-phase could make the temporary depletion of nucleosomes in actively replicated regions more pronounced in the SSDDCS strain

and cause problems in chromatin maturation, leading to differences in later time-points. Another possibility would be that the overall advance in RT and increase in origin firing in early S-phase caused by SSDDCS overexpression could lead to more randomly positioned nucleosomes, which could reduce the number of nucleosomes that are above the threshold of detection using this method.

It is important to mention that these differences were very small (around 1.1 to 1.7% of the total number of nucleosomes for the mid and late time-points) and as such it is unlikely that they would have a major impact on the genome. However, if these differences were taking place in specific regions, such as origins or promoters, they could have a significant impact on the genome homeostasis. In order to address this possibility, nucleosomes were annotated into 4 functional groups based on the following criteria:

- TSS or promoter nucleosomes: nucleosomes located within 350bp upstream and 100bp downstream of the transcription start site (-350bp to +100bp);
- Intragenic nucleosomes: nucleosomes located in coding sequences (100bp downstream of the TSS to the gene 3' end);
- Origin nucleosomes: within origin sequences;
- Intergenic nucleosomes: outside gene bodies, origins or promoters.

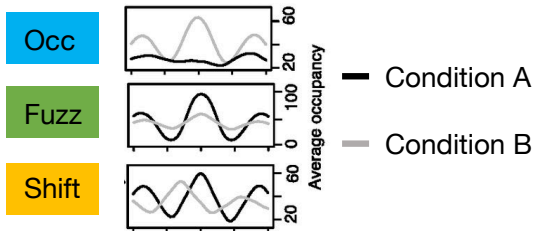
The proportion of different types of nucleosomes was comparable to previous studies[159]: the budding yeast genome is highly compact and mostly occupied by genes, so approximately 70% of all nucleosomes are located in intragenic regions, followed by nucleosomes in promoter regions (~18%), intergenic regions (~12%) and origins (~1%) (Fig. 4.3D-E). Overall, the proportion of different types of nucleosomes was not affected during the time-course or between the 2 strains (Fig. 4.3D-E).

The fact that no changes were observed in the total number and distribution of nucleosomes does not rule out an impact on the chromatin landscape, as the same nucleosome could have slight changes on its positioning or signal between the strains. As such, DANPOS was used to calculate differences between nucleosome positioning and signal at single-nucleotide resolution[159] between the strains (detailed analysis in Methods). DANPOS also allows the classification of dynamic nucleosomes into 3 categories: position shifts, fuzziness and occupancy changes[159] (Fig. 4.4A - right). The exact positions of nucleosomes in each cell in a population might deviate more or less while centred around a most preferred position[159]. This deviation in a cell population is referred to as fuzziness, while occupancy refers to the frequency with which this position is occupied by a nucleosome in the cell population (see Fig. 4.4A cartoon for a visual aid on the different types of dynamic nucleosomes). There is some level of overlap between the different categories, which is explained by the fact that a dynamic nucleosome can have a position shift and an occupancy change simultaneously (this is the reason why the sum of dynamic nucleosomes on each category is higher than the total number of dynamic nucleosomes – Fig. 4.4A - Table).

Using this approach, approximately 2000 to 5000 dynamic nucleosomes were identified in each time-point between the two strains (around 2 to 6% of all nucleosomes). Early time-points including G1 have the highest number of dynamic nucleosomes, suggesting effects which are independent of the advance in replication timing (Fig. 4.4A). For example, of the 4498 nucleosomes classified as dynamic in G1, 813 had a shift in their position (from 20 to 90bp) between the two strains, 1814 had statistically significant fuzziness changes and 4182 had statistically significant occupancy changes (Fig. 4.4A - Table). The cartoon on the right side of figure 4.4A illustrates representative profiles of different types of dynamic nucleosomes between two different experimental conditions.

A

Time-point	Dynamic nucleosomes	Position (20-90bp)	Fuzziness	Occupancy
G1	4498	813	1814	4182
5	4758	942	2084	4467
10	4930	903	2192	4514
15	3753	865	1639	3414
20	3579	1099	1564	3311
25	2994	1104	1638	2703
30	2578	939	1586	2372
35	2152	697	1454	1959
40	1823	607	1271	1651
45	2443	804	1596	2223
50	2714	891	1647	2418
60	2641	786	1642	2392



B

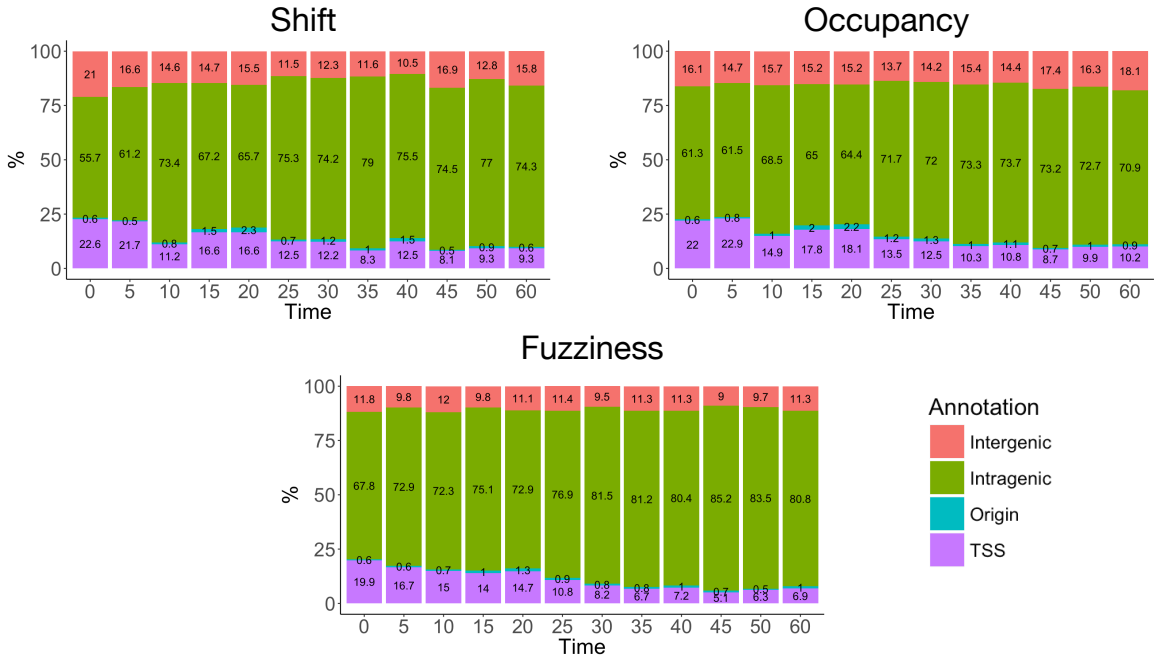


Figure 4.4 – Distribution of dynamic nucleosomes between the two strains during S-phase. A – Total number of dynamic nucleosomes between the two strains on each time-point (DANPOS point_diff_FDR < 0.05). Dynamic nucleosomes were further classified in 3 functional categories. Position shift: 20 to 90bp shift in nucleosome peak position between the two strains; occupancy changes: smt_diff_FDR < 0.05 and fuzziness changes: fuzziness_diff_FDR < 0.05. Cells were colour-coded by column based on the number of different types of dynamic nucleosomes identified in each time-point. Right – cartoon illustrating representative profiles of different types of dynamic nucleosomes between two experimental conditions. **B** – Proportion of dynamic nucleosomes in each time-point annotated to different genomic features as in figure 4.3.

In order to address if any genomic feature was particularly affected, dynamic nucleosomes were then functionally annotated into different genomic features as described previously (see Fig. 4.3 - 4.4B). Interestingly, the proportion of intragenic dynamic nucleosomes increased with S-phase progression for the 3 categories (Fig. 4.4B – green bars). Contrary to what was observed in the gene expression profiles, where most changes seem to be resolved by the end of S-phase, the proportion of intragenic dynamic nucleosomes remained high at later time-points. At 10 and 15

minutes after G1 release there was an increase in the proportion of nucleosomes with position shifts and occupancy changes in origins of replication (Fig. 4.4B – blue bars), which could be a consequence of the overall advance in replication. Strikingly, the proportion of dynamic nucleosomes in promoter regions decreased for all categories during S-phase (Fig. 4.4B violet bars), which could represent significant rearrangements of chromatin during the time-course.

The organisation of nucleosomes in the promoter region plays a major role in gene expression regulation[155], so the proportion of genes with dynamic nucleosomes in promoters that also have an altered transcription profile in the RNA-Seq analysis (Fig. 3.8 – k-means clustering) was analysed. After excluding genes with dynamic nucleosomes in promoter regions in G1 phase, 28% of all genes in the genome had at least one dynamic nucleosome in the promoter region in one or more time-points (Fig. 4.5A – dashed horizontal line). This proportion was comparable between all DE clusters and non-DE genes, except for cluster 1 which had significantly more genes with dynamic nucleosomes in the promoter than expected by chance (Fig. 4.5A).

This analysis was extended for genes with dynamic nucleosomes in the gene body (intragenic), which revealed that 46% of all genes have at least one affected nucleosome in the gene body in one or more time-points (Fig. 4.5B – dashed horizontal line). Cluster 3 and 6 had significantly more genes in this group than expected by chance (Fig. 4.5B). These results indicate that some changes in gene expression could be explained by chromatin changes (or vice versa), and cluster 1 remains a promising group for the identification of links between RT, transcription and chromatin.

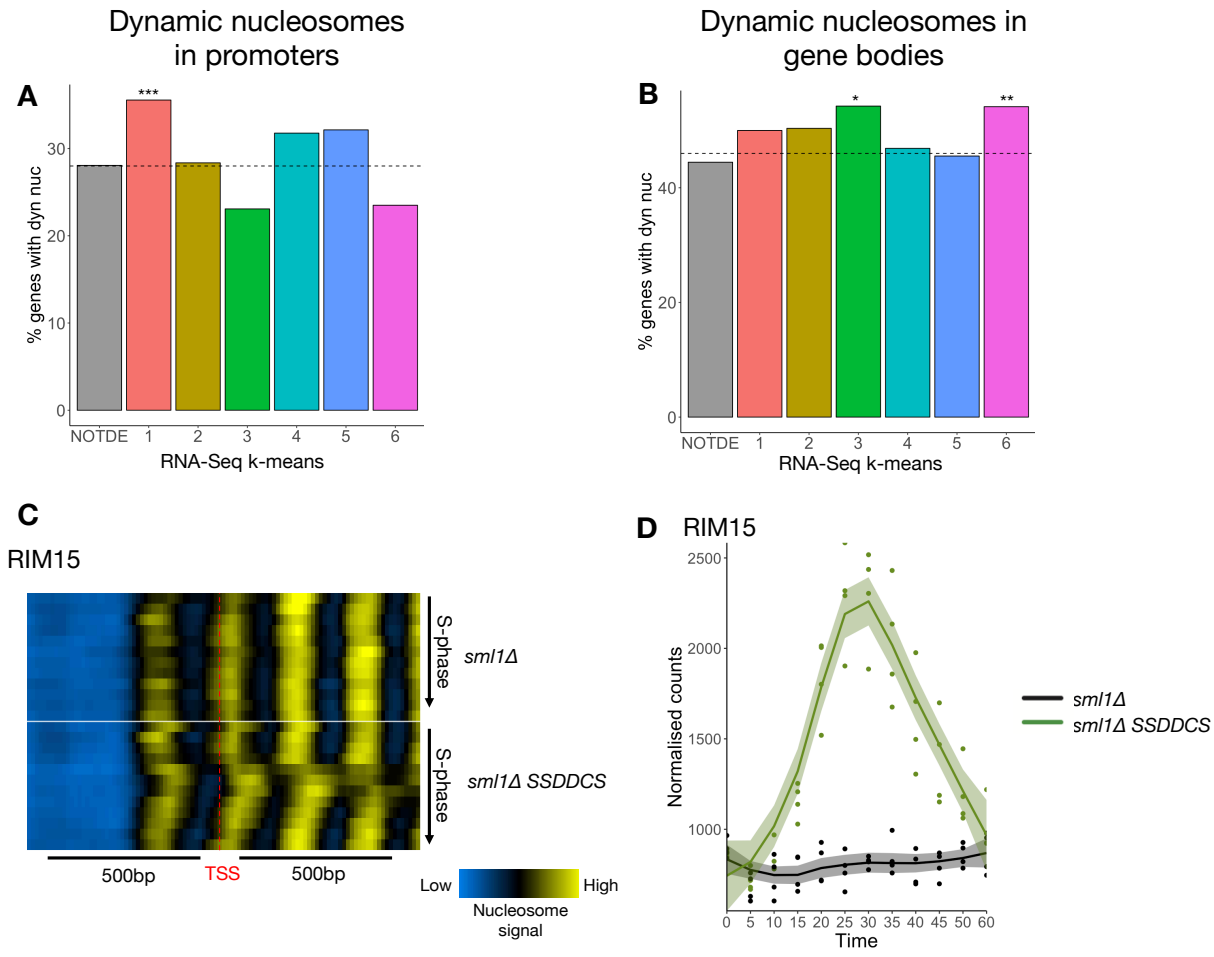


Figure 4.5 – Some DE clusters have more genes with dynamic nucleosomes than expected by chance and there are cases where nucleosome movement associates with gene expression. A – Fraction of genes within each RNA-Seq DE cluster that have dynamic nucleosomes between the two strains in one or more time-points at the promoter region. Genes with dynamic nucleosomes in G1 were excluded. The horizontal dashed line represents the proportion of all genes that have dynamic nucleosome at promoters in one or more time-points (28%). p-values are from an exact binomial test. **B** – Same as A but for intragenic nucleosomes. p-values are from an exact binomial test, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. **C** – Nucleosome heatmap centred around the TSS of *RIM15*, a meiotic regulator from cluster 1. The vertical red dashed line marks the TSS and each row represents one time-point, from G1 to 60 from top to bottom, respectively. The two strains are separated by a white horizontal line and the heatmap is coloured according to the density of MNase-Seq reads: yellow – high density of reads corresponding to nucleosome peaks, blue – low density of reads, corresponding to nucleosome depleted regions. **D** – Gene expression profile of *RIM15*, represented as previously described, allowing a direct comparison with the nucleosome profile.

A closer look at the nucleosome landscape of genes from cluster 1, such as the meiotic regulator *RIM15*, allowed the identification of events where movement of the +1 nucleosome, accompanied by overall disorganisation of chromatin in the gene body, is associated with changes in gene expression (Fig. 4.5C-D). Strikingly, for

this particular gene, the nucleosome shift at the promoter and loss of organisation of nucleosomes in the gene body took place during the time-points of maximum expression. Moreover, nucleosome organisation in this locus was re-established by the end of time-course, which was accompanied by a re-establishment of the wild-type gene expression levels (Fig. 4.5C-D).

It should be mentioned at this stage that these results also indicate that many of the observed effects on gene expression are independent of chromatin changes in the SSDDCS strain. Moreover, in some cases chromatin changes in the SSDDCS strain do not result in changes in gene expression (there are many genes which are not DE that also have dynamic nucleosomes – Fig. 4.5A-B). Similar to what was observed in the RNA-Seq analysis, the identification of changes per time-point might mask more dynamic patterns during S-phase. Rather than limiting the analysis to pairwise comparisons between time-points, more robust approaches were implemented in order to better understand the effect of advancing replication timing on the chromatin landscape and nucleosome patterns.

4.2 - Genome-wide advance in replication timing affects chromatin conformation in promoters and gene bodies

Given the role of the +1 nucleosome on gene expression regulation, the impact of the overall advance in RT on the movement of the +1 nucleosome throughout S-phase was analysed, focusing on the genes with altered transcription profiles from the RNA-seq analysis as done previously. To identify the +1 nucleosome, the following approach described in Nocetti and Whitehouse (2016) was used: nucleosomes assigned to the 100bp window around the TSS ranging from 20bp upstream to 80bp downstream (-20bp to 80bp)[155] were considered to be the +1 nucleosome. Consistent with the Nocetti and Whitehouse data, 3551 genes with a nucleosome present on this window were identified.

To analyse the movement of the +1 nucleosome during S-phase, the +1 position was normalised relative to the TSS (relative position = +1 nucleosome position – TSS position) and plotted as a heatmap (Fig 4.6A). Strikingly, the +1 nucleosome tends to move further into the gene body in the SSDDCS strain compared to the *sml1Δ* strain. This shift took place in early to mid S-phase time-points, but by the end of the time-course the +1 nucleosome was back to a more upstream position, closer to the TSS (Fig. 4.6A). This effect at > 3000 genes is similar to what was observed at the *RIM15* locus (Fig. 4.5C).

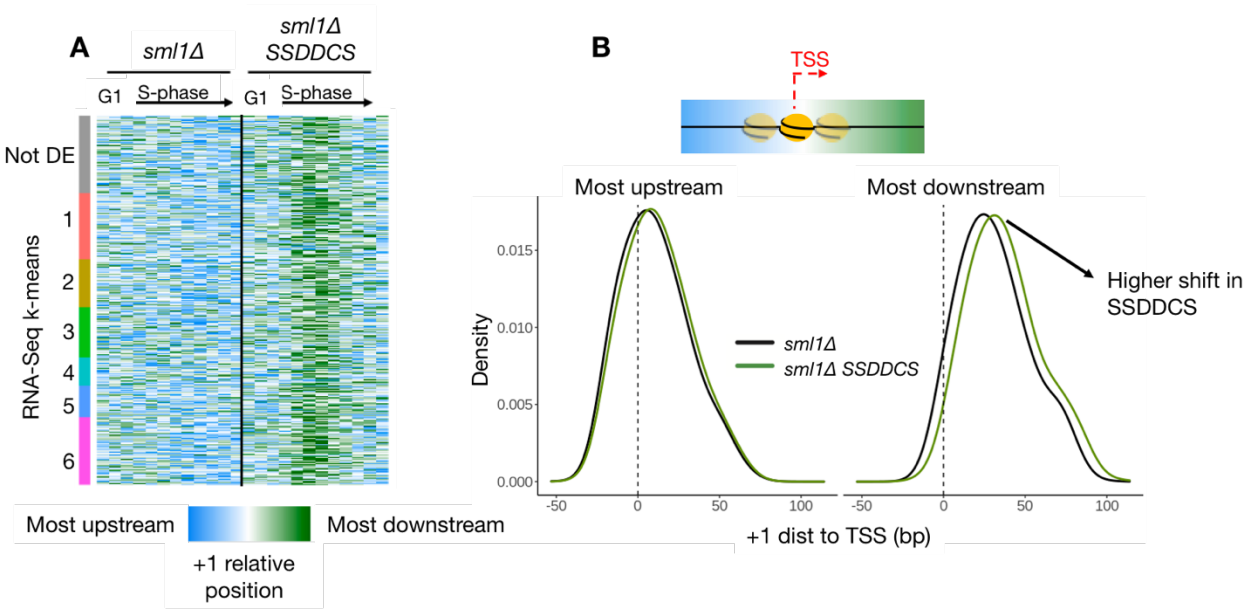


Figure 4.6 – Advance in replication timing affects the positioning of the +1 nucleosome genome-wide. A – Position of +1 nucleosome relative to TSS (+1 position – TSS) represented as a heatmap for the 940 DE genes plus 250 random non-DE genes with an annotated +1 nucleosome. Blue time-points represent the time-points at which the +1 nucleosome is at the most upstream position (close to TSS), while green represent the time-points at which it is at the most downstream position (into the gene body). Data is normalised by row as done for previous heatmaps. Each row corresponds to a single gene and each column to a single time-point (G1 to 60 from left to right, respectively). The two strains are separated by a black vertical line. The colour key on the left of the heatmap allows the annotation of each gene to its corresponding DE cluster. Gray – 250 random non-DE genes. **B** – Distribution of +1 nucleosome relative position for all genes on its most upstream (left) and most downstream (right) position, represented as a density plot. Dashed vertical line represents the TSS location. On its most upstream position, the +1 nucleosome is close to the TSS and this position is similar between the two strains, while on its most downstream position it is further into the gene body in the SSDDCS compared to the *sml1Δ*. The cartoon on the top uses the colours from the heatmap in order to be used as a visual guide to interpret the plots.

Comparison of the distribution of the most upstream versus the most downstream position of the +1 nucleosome for all genes where this nucleosome was identified, further confirmed the higher genome-wide shift upon SSDDCS overexpression (Fig. 4.6B). The most upstream position of the +1 nucleosome was close to the TSS, and similar between the two strains (Fig. 4.6B left). On the other hand, the most downstream position of the +1 nucleosome was further away from the TSS in the SSDDCS compared to the *sml1Δ* strain (Fig. 4.6B right). The heatmap suggests that the higher shift of +1 nucleosome into the gene body is a genome-wide effect, rather than an effect specific to DE genes (Fig. 4.6A).

At its most downstream position, the higher shift of +1 nucleosome into the gene body was observed for all groups of genes, suggesting a global effect on nucleosome positioning upon SSDDCS overexpression (Fig 4.7A). Despite this overall effect, some groups of DE genes showed a slightly more pronounced shift, such as cluster 1 and 4 (Fig. 4.7A). These results demonstrate the higher shift of the +1 nucleosome movement into the gene body upon SSDDCS overexpression by comparing the most extreme positions, but do not address the degree of +1 mobility during the time-course.

To analyse the overall mobility of the +1 nucleosome during S-phase, the standard deviation of its position relative to the TSS (determined above) was calculated for each +1 nucleosome: higher standard deviations correspond to more mobile nucleosomes, while lower standard deviations correspond to static nucleosomes. Comparison of the +1 mobility between the two strains showed that, overall, the +1 nucleosome positioning was more dynamic in the SSDDCS for all groups of genes, including non-DE (Fig. 4.7B). Moreover, the difference between the standard deviation calculated for each strain (SSDDCS - *sml1Δ*) provides a quantitative assessment of the differences in mobility, as follows: if the difference is equal to 0 the +1 nucleosome has the same degree of mobility in both strains while if the difference is below or higher than 0 it means that the +1 nucleosome is more mobile in the SSDDCS or *sml1Δ*, respectively (Fig. 4.7B).

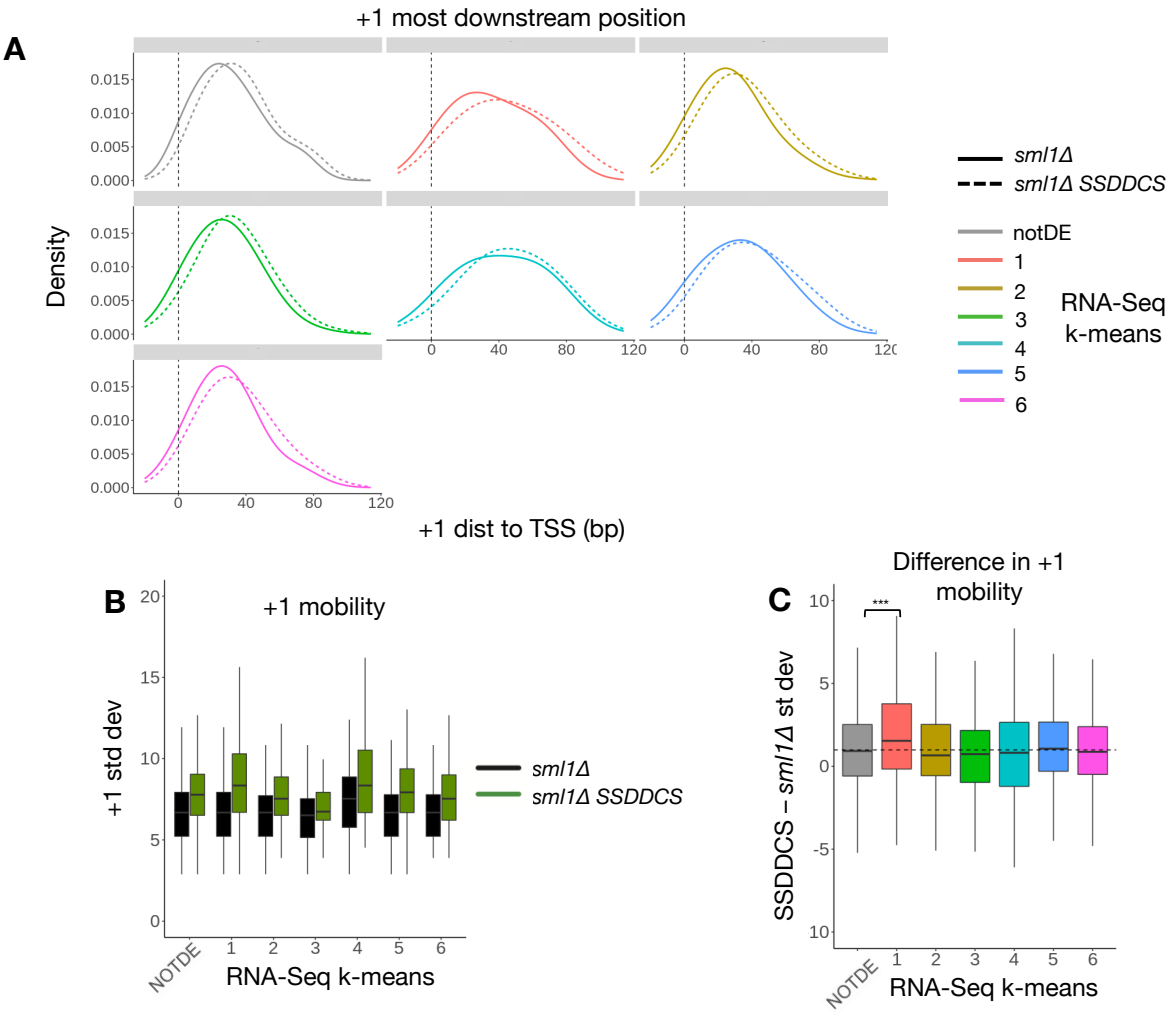


Figure 4.7 – The +1 nucleosome is more mobile upon SSDDCS overexpression, with maximum mobility on cluster 1 genes. A – Distribution of most downstream position of +1 nucleosome for all genes split by RNA-Seq cluster, represented as a density plot. Dashed vertical lines mark the TSS. Each strain is represented by lines with different shapes. The further advance into the gene body was present in all groups, but the difference between the strains were slightly higher in cluster 1 and 4. **B** – Distribution of +1 mobility during S-phase, represented as the standard deviation of relative positions throughout the time-course. Overall, the +1 nucleosome was more mobile in the SSDDCS strain in every group of genes. **C** – Distribution of the difference in +1 mobility between the 2 strains (standard deviation SSDDCS – standard deviation *sml1Δ*) for the different group of genes. Horizontal dashed line marks the median difference for all genes in the genome. Cluster 1 is the only group which is significantly different from the genome median. p-value is from pairwise comparisons of k-means cluster 1 versus the non-DE genes using Wilcoxon rank sum test, *** $p < 0.001$.

As expected, all groups of DE genes as well as the non-DE exhibited an average difference in standard deviation higher than 0 (Fig. 4.7C), further confirming that the +1 nucleosome is more mobile upon SSDDCS genome-wide (Fig. 4.7B). Strikingly, comparison of the difference in standard deviation (i.e. in +1 mobility) between the two strains showed that cluster 1 genes were the only group significantly different from the non-DE, suggesting higher +1 mobility in this group of genes upon SSDDCS overexpression (Fig. 4.7C). These analyses illustrate the increase in +1 mobility upon SSDDCS overexpression, which is consistent across all genes (Fig. 4.6 and 4.7) but slightly higher in cluster 1 (Fig. 4.7C).

To analyse nucleosome organisation in the gene body during S-phase, the following approach was used: the autocorrelation function (ACF) was used to determine the pattern of organisation of the first 4 nucleosomes as a proxy for chromatin organisation, as described in Gutiérrez et al.[160]. Genes with higher ACF values have well-phased and organised nucleosomes, while lower values represent poorly organised chromatin (Fig. 4.8A). ACF values were calculated for all genes in each S-phase time-point and plotted as a heatmap (Fig. 4.8B). This analysis was restricted to genes > 700bp (as 4 nucleosomes are needed for ACF calculation), which corresponds to approximately 70% of all genes.

During S-phase progression, there is a transient decrease in nucleosome organisation, most likely caused by the passage of replication forks, but once replication is completed the nucleosome organisation is re-established (Fig. 4.8B-C). Strikingly, a greater decrease in nucleosome organisation was observed upon SSDDCS overexpression both in single genes (Fig. 4.8C – *RIM15*) and genome-wide (Fig. 4.8B). Comparison of the ACF values of *RIM15* in each time-point illustrates the strength of the analytical approach used to identify meaningful chromatin perturbations (compare heatmap from Fig. 4.5C and Fig. 4.8C). In the *sml1Δ* strain, nucleosomes in the *RIM15* gene body were well positioned and static throughout S-phase, hence the small variation in ACF values. On the other hand, the shift of the +1 nucleosome and overall disorganisation of the first nucleosomes into the gene body (Fig. 4.5C) caused the observed drop in ACF values in mid S-phase (Fig. 4.8C).

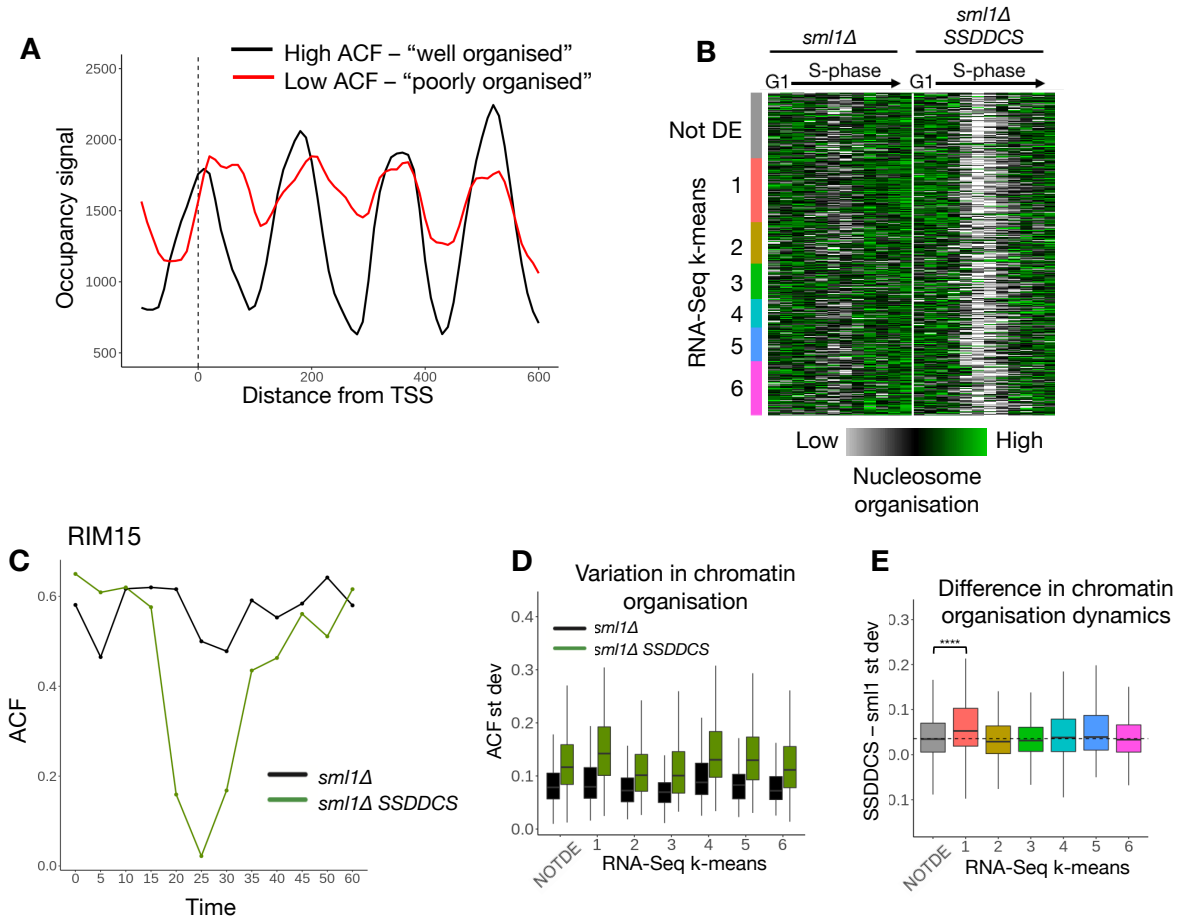


Figure 4.8 – Genome-wide advance in replication timing causes a drop in chromatin organisation in gene bodies during S-phase, with maximum differences on cluster 1 genes. A – Nucleosome occupancy signal of the first 4 nucleosomes after the TSS. Two extreme examples were selected to illustrate the autocorrelation function (ACF) analysis: nucleosomes in black are well phased (i.e. well defined peak and valleys) and well positioned (similar inter-nucleosome distance), corresponding to high ACF values. On the other hand, nucleosomes in red are poorly phased and disorganised, resulting in low ACF. ACF values were calculated for every gene in every time-point. Vertical dashed line marks the TSS. **B** – Heatmap of ACF values as a proxy of nucleosome organisation during S-phase for the 4537 genes longer than 700bp. Each row corresponds to a single gene and each column to a single time-point (G1 to 60 from left to right, respectively). The two strains are separated by a white vertical line. A colour key was added on the left of the heatmap to annotate gene to its corresponding DE cluster. Gray – 350 random non-DE genes. White and green time-points represent the time-points at which the chromatin is mostly disorganised and organised, respectively. **C** – ACF values in *RIM15* locus illustrate the sharp drop in organisation in the SSDDCS strain during mid S-phase. **D** – Distribution of organisation dynamics during S-phase, represented as the standard deviation of ACF values during the time-course. This value was higher upon SSDDCS overexpression in every group of genes. **E** - Distribution of the difference in ACF standard deviation (stdev SSDDCS – stdev *sml1⊿*) for the different group of genes. Horizontal dashed line marks the median difference for all genes in the genome. Cluster 1 is the only which is significantly different from the genome median. p-value is from pairwise comparisons of k-means cluster 1 versus the non-DE genes using Wilcoxon rank sum test, **** $p < 0.0001$.

Despite the more pronounced disorganisation of chromatin upon SSDDCS overexpression in mid S-phase, nucleosome organisation was re-established by the end of the time-course, suggesting that cells were able to resolve this transient effect (Fig. 4.8B-C). Once again, to address whether any particular cluster of DE genes was significantly more affected, the standard deviation of ACF values during S-phase was calculated for each gene in each strain (higher standard deviations would be a result of a stronger variation in organisation during the time-course and vice-versa). As expected from the heatmap (Fig. 4.8B), all clusters of genes showed higher disorganisation in the SSDDCS strain compared to the *sml1Δ,* as illustrated by the higher average standard deviations (Fig. 4.8D). The difference between standard deviation in each strain allowed a quantitative assessment of the differences in chromatin organisation between the two conditions, as done previously for the +1 mobility (Fig. 4.7C). Strikingly, cluster 1 was the only group where the difference between ACF variation (SSDDCS - *sml1Δ*) was significantly different from the genome average (Fig. 4.8E), further confirming that despite the genome-wide trend, these genes represent the most affected group.

The distinct analytical approaches represented by the heatmaps in figure 4.6A and figure 4.8B illustrate a similar trend, suggesting that the overall loss of chromatin organisation could be a consequence of the +1 nucleosome movement. Indeed, a significant shift of the +1 nucleosome during mid S-phase as well as a drop in nucleosome organisation was observed in *RIM15* (Fig. 4.5C and Fig. 4.8C). These results show a transient genome-wide effect on chromatin organisation, which is restricted to mid S-phase as chromatin organisation is re-established in later time-points. The higher rates of replication initiation in early S-phase, caused by the overexpression of SSDDCS, could have an impact on nucleosome deposition and chromatin maturation and as S-phase progresses and initiation rates decrease cells might be able to re-establish normal chromatin organisation (see Discussion).

Considering that cluster 1 genes were up-regulated upon SSDDCS overexpression and that these genes showed the greatest changes in chromatin organisation at promoters and gene bodies during the time-points of maximum expression makes them a promising set of genes for downstream analysis and identification of links with RT. Still, the genome-wide effect on the chromatin landscape upon the

advance in RT, both at DE and non-differentially expressed genes, implies that chromatin disorganisation in promoters and gene bodies is not sufficient to cause changes in gene expression. On the other hand, the fact that non-DE genes also have changes in chromatin suggests that some of the chromatin changes observed are not simply due to altered transcription (which could be the case for some DE genes).

4.3 - Genome-wide advance in replication timing affects chromatin conformation on transcription-factor binding sites

The fact that some DE clusters share common biological functions, suggests that they are regulated by common mechanisms, namely common transcription-factors (TFs). Differential TF binding events could help explain the differences observed in gene expression and chromatin upon SSDDCS overexpression: an impact of RT on the chromatin landscape could lead to the binding of TFs to cryptic sites or sites which are unavailable for binding during a normal S-phase, or perturb the homeostasis of TFs which are regulated during the cell cycle (these two hypothesis will be discussed in more detail in subsequent chapters). Moreover, perturbation of TF binding dynamics could also be the cause of changes in chromatin landscape, through a positive feedback loop where perturbed chromatin would allow TF binding events which could cause further chromatin effects.

In order to identify mechanisms that contribute to the differential expression of genes after a global advance in replication timing, the list of DE genes identified was compared with published TF binding datasets. For the reasons already described in previous sections, the analysis was focused on cluster 1. The Yeast Transcription Factor Specificity Compendium (YetFaSCo) is a database of all available budding yeast TFs, comprising information from 133 TF binding studies analysing 256 DNA binding proteins[161]. One of the features of this database is a search tool that allows the identification of regulators which are enriched for particular groups of genes, by using the database to find how well each TF correlates with the query data (i.e. the list of genes provided). This search is performed using the rank sum test, which

tests whether the list of genes provided is significantly different from the rest of the genome.

The two outputs from the rank sum test are a p-value representing how significant the association is and the area under the receiving operator characteristic (ROC) curve, which represents whether the list of genes provided is significantly enriched (ROC > 0.5) or depleted (ROC < 0.5) for targets which are bound by each TF present on the database[161]. The rank sum test was used to identify which TFs were significantly enriched for binding of genes from cluster 1, and the results were plotted as a volcano plot (Fig. 4.9A). Overlaps were considered significant if the p-value was lower than 0.05. Each data point corresponds to data from a single ChIP study, so the same TFs can appear more than once if it was analysed by more than one study. Using this analysis, 33 TFs were identified as significantly enriched for the binding of cluster 1 genes, while 124 were significantly depleted (Fig. 4.9A). The identification of more TFs significantly depleted is expected, as a relatively small group of genes (480 genes) is being compared against the whole genome, so the targets of most TFs won't be present in this list. The TFs enriched for cluster 1 binding belonged to 3 broad categories: stress response, meiosis control and chromatin regulation (Fig. 4.9A). This analysis agrees with the GO enrichment (Fig. 3.11), which identified a significant enrichment for genes involved in meiosis and sporulation.

As described in previous sections, the meiotic gene expression should be off in mitotic cells and the up-regulation of these genes upon perturbations of RT draws an interesting parallel with metazoans, where RT changes are associated with cellular fate transitions (3.12A). Meiotic gene expression is silenced during mitosis by the repressor protein Ume6, which binds the URS1 consensus sequence[173] (Fig. 4.9C) and recruits the histone deacetylases Rpd3 and Sin3 to the promoter regions of early meiotic genes (EMG) (Fig. 4.9B), resulting in repression of these genes by hypoacetylation[174]. Ume6 was identified in a screen for haploids that express meiotic genes during mitosis[175], and it has been shown that its degradation in early meiosis leads to activation of meiotic gene expression[176]. Ume6 was one of the hits identified as enriched for the binding of genes from cluster 1 (Fig. 4.9A). Therefore, one possibility for the activation of meiotic genes in the SSDDCS strain is the

disruption of Ume6 binding dynamics (Fig. 4.9B), which could be a consequence of an impact on chromatin organisation in these regions.

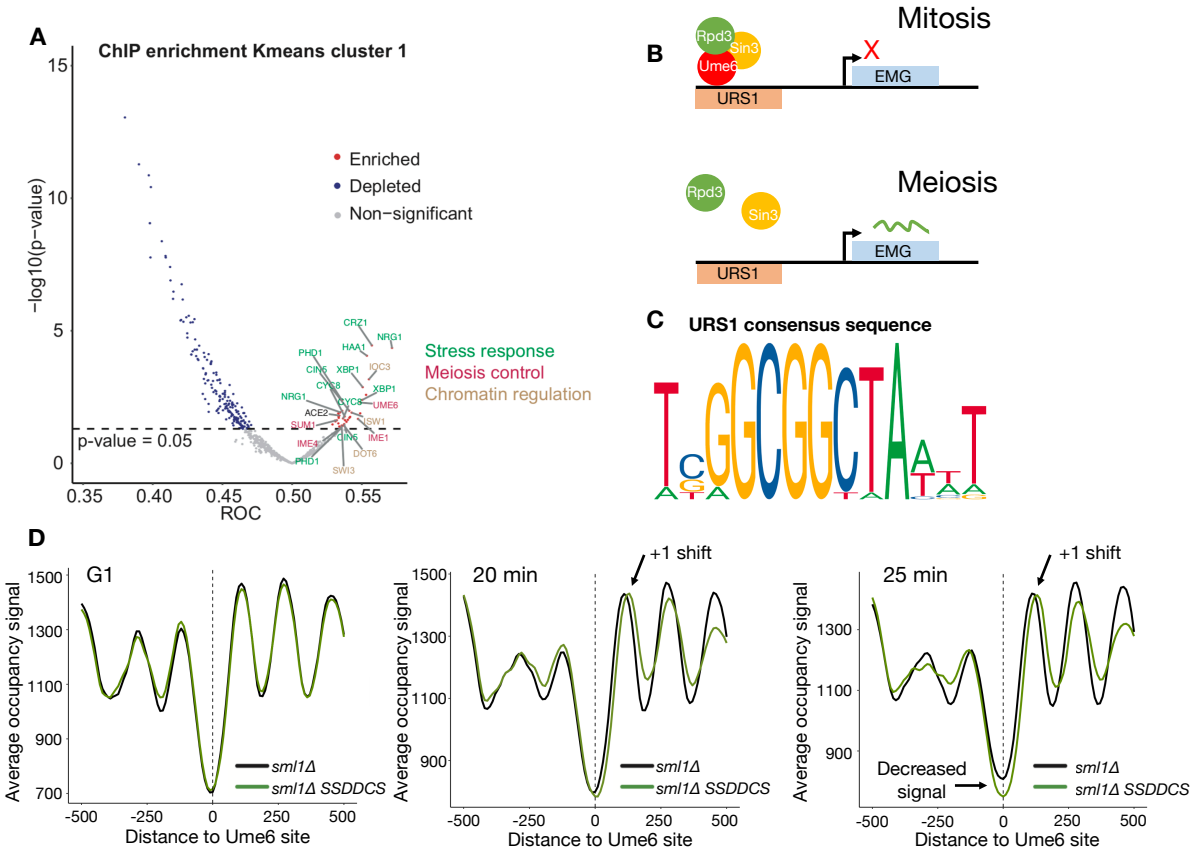


Figure 4.9 – Known regulators of the meiotic gene expression programme, such as Ume6, are enriched for binding of cluster 1 genes and the nucleosome landscape is affected in these regions upon SSDDCS overexpression. A – Output of YeTFaSCo rank sum test for cluster 1 genes plotted as a volcano plot. p-value was log-normalised and the dashed horizontal line sets the minimum value required for a TF to be considered significantly associated with cluster 1 genes (p-value = 0.05). ROC stands for the area under the receiving operator curve. TFs which are significantly enriched or depleted for cluster 1 binding are coloured in red and blue respectively (ROC > 0.5 or ROC < 0.5), while not significant hits are coloured in grey (p-value > 0.05). Enriched TFs were coloured according to functional categories. **B** – Repression of meiotic gene expression is mediated by Ume6. During mitosis, Ume6 binds to URS1 consensus sequence in the promoter region of early meiotic genes (EMG) and blocks gene expression by recruiting the histone deacetylase proteins Rpd3 and Sin3. Upon environmental signals for meiotic onset, Ume6 is degraded allowing the expression of these genes. **C** – URS1 consensus sequence represented as a sequence logo. For each position, the stack of letters represent the frequency of each nucleotide (the total height of each letter on each position represents the degree of conservation). Positions with more than one nucleotide are ordered from most to least frequent, from top to bottom. **D** – Average nucleosome profiles around the 89 URS1 locations identified in promoters of cluster 1 genes in G1, 20 and 25 minutes. While the profiles were identical in G1, there was a decrease in the MNase-Seq signal at the Ume6 binding site in the SSDDCS strain accompanied by a +1 nucleosome shift. Vertical dashed line marks the URS1 sequence, which corresponds to the Ume6 binding site.

To address this possibility, the URS1 consensus sequence (Fig. 4.9C) was extracted from the JASPAR database[162] and the locations of all URS1 sequences in the budding yeast genome were identified using Find Individual Motif Occurrences (FIMO)[163] (detailed analysis in Methods). Of the 2875 URS1 sequences identified genome-wide, 89 were located 1kb upstream of the TSS of genes from cluster 1. Analysis of the average nucleosome profile centred around these 89 locations, showed a differential MNase-Seq signal in the Ume6 binding site, accompanied by movement of the +1 nucleosome during S-phase, but not in G1 cells (Fig. 4.9D). This suggests a potential role of replication timing for TF binding and chromatin organisation at these locations.

To analyse the chromatin organisation around Ume6 binding sites and corresponding promoter regions in single genes, heatmaps of *IME2* and *NDT80* spanning the Ume6 binding sites were generated. Using the approach described above for the identification of Ume6 binding sites genome-wide, two Ume6 binding sites were identified on both *NDT80* and *IME2* promoter regions. *NDT80* has one site 164bp and another 290bp upstream of the TSS while *IME2* has one site 451bp and another 546bp upstream of the TSS (Fig. 4.10 – vertical solid red lines). The fact that these two sites are nearly equidistant in both genes (126bp and 95bp for *NDT80* and *IME2*, respectively), suggests that the simultaneous presence of these two sites might be needed for proper Ume6 binding and regulation of the target genes. Strikingly, chromatin conformation was affected in both genes nearby to the Ume6 binding sites upon overexpression of SSDDCS (Fig. 4.10).

The two Ume6 sites in the *NDT80* promoter are located in a nucleosome depleted region next to two well-positioned nucleosomes. While these nucleosomes were mostly static in the *sml1Δ* strain, a slight "opening" of these nucleosomes was observed in the SSDDCS strain, which increased the length of the nucleosome depleted regions transiently (Fig. 4.10 – Left).

In the *IME2* locus, the Ume6 sites are located further upstream of the TSS, nearby a well-positioned nucleosome. A movement of this nucleosome was observed upon SSDDCS overexpression, while in the *sml1Δ* strain this nucleosome was mostly static (Fig. 4.10 – Right). As seen for most of the effects caused by SSDDCS overexpression, the nucleosome positioning was re-established by the end of the time-course.

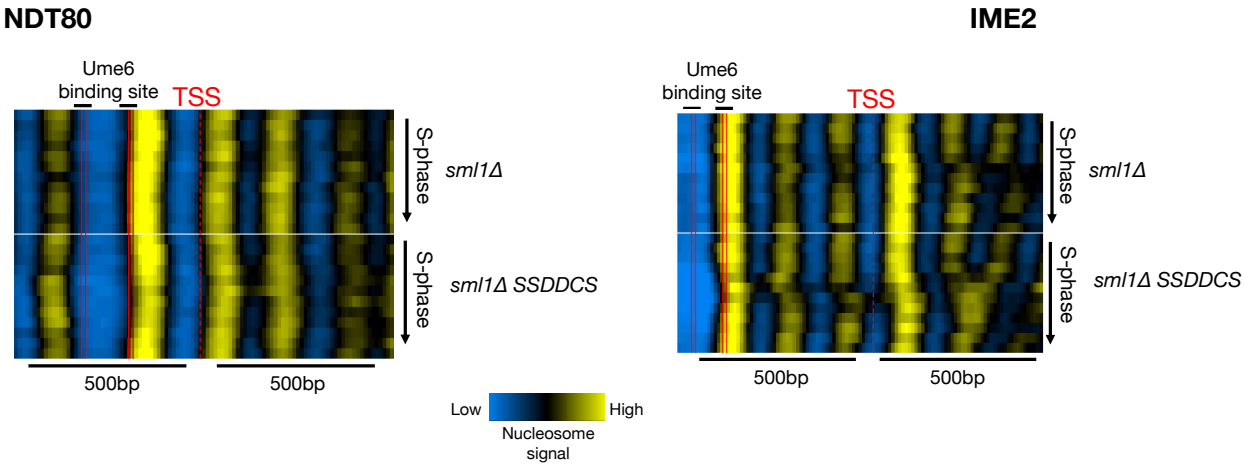


Figure 4.10 – Chromatin conformation of *NDT80* and *IME2* is affected in Ume6 binding sites upon SSDDCS overexpression. Nucleosome heatmaps of *NDT80* (left) and *IME2* (right) loci spanning the TSS and Ume6 binding sites. Each of these genes has two Ume6 binding sites (vertical solid red lines). The vertical dashed red line marks the TSS and each row represents one time-point, from G1 to 60 from top to bottom, respectively. The two strains are separated by a white horizontal line and the heatmap is coloured according to the density of MNase-Seq reads: yellow – high density of reads corresponding to nucleosome peaks, blue – low density of reads, corresponding to nucleosome depleted regions. Nucleosomes located close to the Ume6 binding sites were affected upon SSDDCS overexpression.

Both genes have different effects on the chromatin conformation downstream of the promoter region: *NDT80* has a slight shift of +1 nucleosome (Fig. 4.10 – Left), while for *IME2* a dramatic movement of the nucleosomes in the gene body was observed, including what seems to be two nucleosomes merging into one in the SSDDCS strain (Fig. 4.10 – Right).

In mitotic haploid cells, Ume6 should be bound to the promoter of these two genes to repress their expression. Upon overexpression of SSDDCS these genes become activated (Fig. 3.12B) and the work presented in this thesis has shown that this effect is dependent on advanced replication of these loci (origin deletion

experiments in Fig. 3.13). The genome-wide advance in RT could be affecting the chromatin landscape in a way that blocks the correct binding of Ume6, thus allowing the untimely expression of these genes. As replication is completed, chromatin is re-established and Ume6 might possibly re-bind to these locations, which would explain why the expression of these genes goes back to G1 levels by the end of the time-course (Fig. 3.12B). Moreover, Ume6 recruits other proteins that also have roles on chromatin regulation, which could cause further alterations in chromatin organisation. However, it is also possible that the chromatin changes observed in these two genes are simply due to the transcriptional activation upon SSDDCS overexpression.

Some of the TFs identified as significantly enriched for binding of cluster 1 genes have roles as general repressors of gene expression such as Cyc8 (Fig. 4.9A) which forms a complex with Tup1 and is involved in the formation of heterochromatin in sub-telomeric regions[177] and in the regulation of expression of more than 300 genes, through the recruitment of this complex by different TFs to their target promoters[178]. Some of these TFs are also hits of the ChIP enrichment analysis, such as Phd1 and Nrg1 (Fig. 4.9A). These results suggest that a complex series of events could be taking place upon SSDDCS overexpression and global advance in RT, including loss of heterochromatin formation and perturbations of transcription-factor binding dynamics during S-phase.

These results suggests that replication timing could have evolved to ensure the correct TF binding dynamics during S-phase and during cellular fate transitions. In order to explore this possibility and link the observations from RT, gene expression and chromatin organisation analyses, I decided to analyse TF binding genome-wide upon changes in RT.

Chapter 5 – Effect of a perturbed replication timing programme on transcription-factor binding dynamics

5.1 - MNase-Seq as a tool to study transcription-factor binding dynamics genome-wide

The observations that a genome-wide advance in RT affects chromatin conformation in promoters and gene bodies, but also in TF binding sites (chapter 4), suggest that the SSDDCS overexpression could perturb TF binding dynamics during S-phase. In order to analyse the effect of a global advance of RT on TF binding genome-wide in an unbiased way, the SSDDCS overexpression system was used in combination with an optimised version of MNase-Seq that allows the isolation of fragments below the size of single nucleosomes, which correspond to the footprint of DNA binding proteins such as TFs (Fig. 5.1A-B).

This protocol was based on a study from Gutiérrez et al., which used MNase-Seq in combination with EdU labelling to analyse TF binding and chromatin maturation dynamics behind the replication fork in budding yeast[160]. Digestion of chromatin with a lower concentration of MNase for a longer period of time at a lower temperature (5U of MNase for 20 minutes at room temperature compared to 90U for 3 minutes at 37°C used for the isolation of mono-nucleosomes (Fig. 4.1) – see Methods for detailed protocol), allowed the isolation of sub-nucleosomal sized fragments (Fig. 5.1C).

Using this optimised protocol, most fragments still correspond to mono, di and tri-nucleosomes, while the sub-nucleosomal population represents a small fraction of the isolated DNA (Fig. 5.1C, dotted square). Attempts to further digest the nucleosomal populations could lead to the digestion of the sub-nucleosomal fraction because these fragments are preferentially digested by MNase. Despite the abundance of fragments in the nucleosome fraction, smaller fragments are preferentially amplified during the PCR step of library preparation. As a result, most of the sequenced fragments were smaller than 150bp, which correspond to the

footprint of DNA binding proteins, such as TFs (Fig. 5.1D). This approach was named subMNase-Seq.

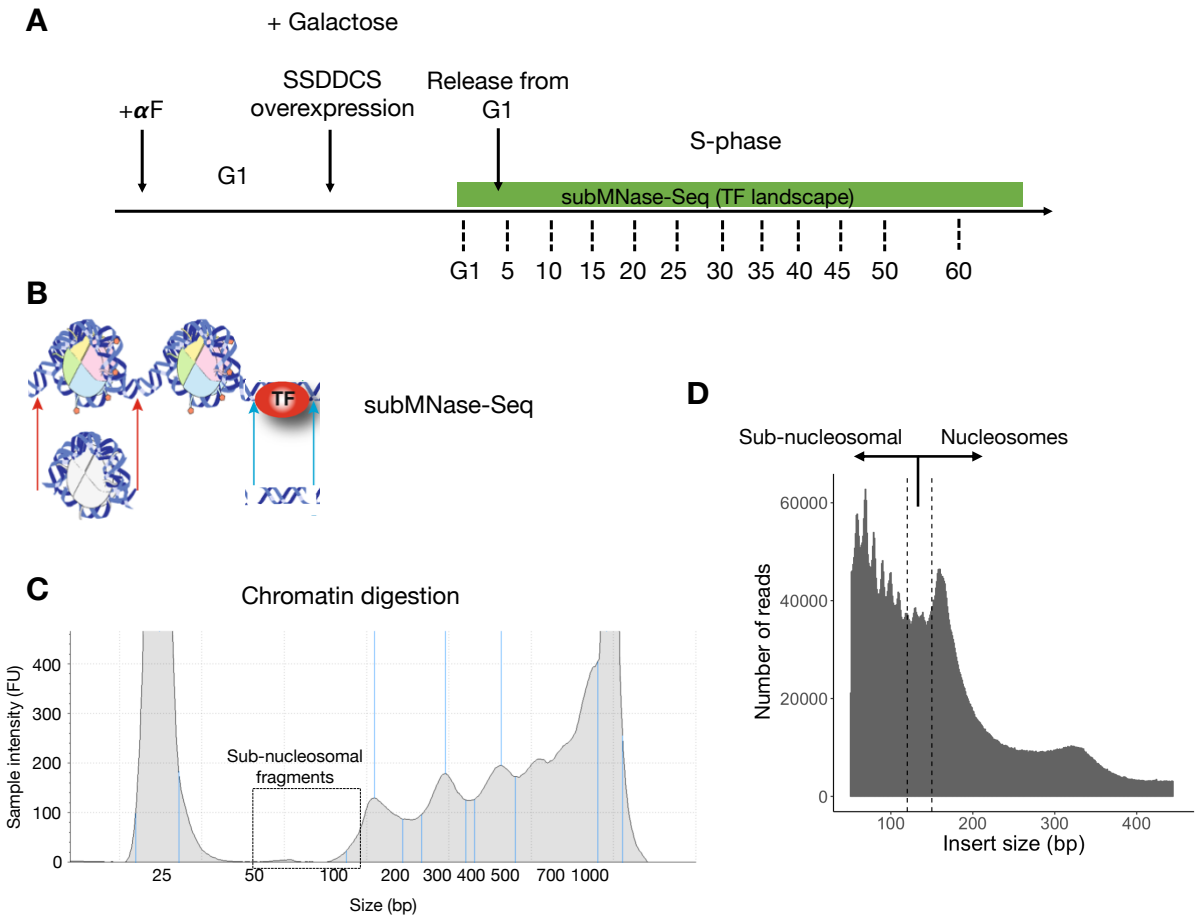


Figure 5.1 – subMNase-Seq allows the isolation of sub-nucleosomal fragments, corresponding to transcription-factor binding events. A - Experimental system described in previous sections, where a cell population overexpressing the SSDDCS is released into a synchronous S-phase. Samples were collected every 5 minutes for subMNase-Seq. **B** - Micrococcal nuclease (MNase) digests naked DNA and optimisation of digestion conditions allows the isolation of fragments with different sizes, corresponding to nucleosome positions and TF binding events. **C** - Agilent D1000 ScreenTape quantification of one DNA sample after MNase digestion optimised to isolate sub-nucleosomal sized fragments. Mono (~170bp), di (~340bp) and tri-nucleosomes (~530bp) represent the majority of the isolated fragments, but a small population of sub-nucleosomal events (dotted square) is also present. Chromatin was digested with 5U of MNase for 20 minutes at room temperature. The abundant populations at 25bp and >1000bp correspond to the lower and upper marker respectively, used to determine the size of the isolated fragments. **D** – Distribution of number of sequenced reads per insert size for one of the samples after MNase digestion. Dashed vertical lines mark 120 and 150bp. During the PCR step of library preparation, smaller fragments are preferentially amplified, so most sequenced fragments were below 120bp, which includes TF binding events. As expected, there was still a considerable amount of mono-nucleosome sized fragments, which have to be kept to avoid losing sub-nucleosomal fragments. This distribution was similar across all samples.

To confirm that this approach isolated genuine TF binding locations, the GAL1-10 promoter was analysed in detail. The SSDDCS strain has a second copy of Sld2, Sld3, Dbf4, Dpb11, Cdc45 and Sld7 under the control of the GAL1-10 promoter[76]. This promoter is bidirectional, so the same GAL1-10 sequence can regulate the expression of 2 limiting factors simultaneously (Fig. 5.2A). As a result, it is expected that the SSDDCS samples will have more reads mapping to the GAL1-10 reference sequence compared to the *sml1Δ* strain (Fig. 5.2A).

As expected, the read coverage around the Gal80 binding sites, which is part of the galactose regulatory network that binds the GAL1-10 promoter[179], was significantly higher in the SSDDCS compared to the *sml1Δ* strain (Fig. 5.2B – G1 time-point). This increase in signal in the GAL1-10 native promoter is consistent throughout the entire time-course and is not due to differences in sequencing yield, as seen from the surrounding sequences (Fig. 5.2B). This result confirmed that genuine TF binding events can be identified using this method.

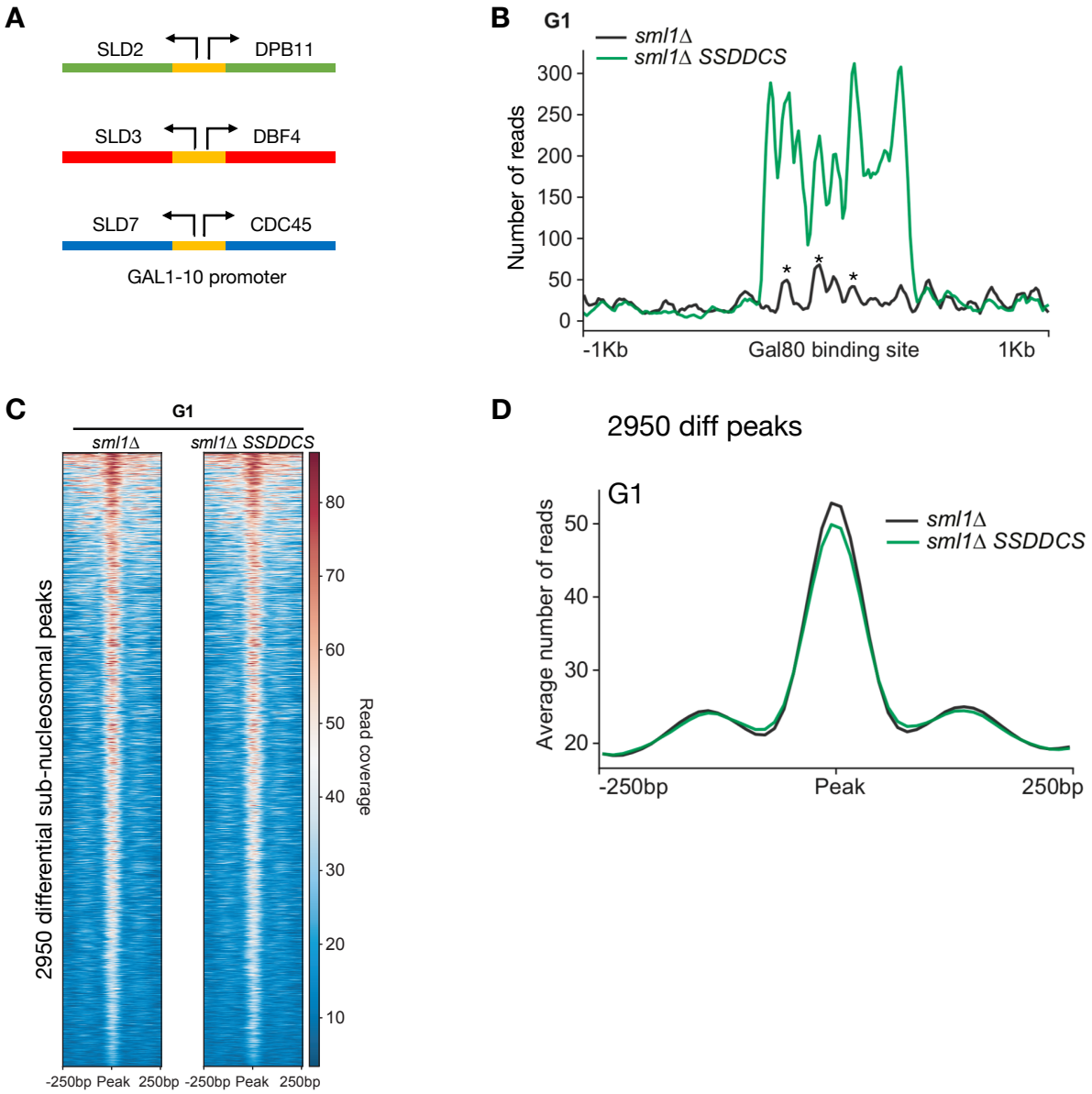


Figure 5.2 – Increased signal in the GAL1-10 locus confirms the overexpression of the SSDDCS and the utility of this approach to identify TF binding events genome-wide. A – Design of SSDDCS strain. The six limiting factors were cloned in the genome as second copies, under the control of the GAL1-10 bidirectional promoter. **B** – Read coverage around the GAL1-10 promoter (Gal80 binding site) in G1. As expected, the signal is significantly higher in the SSDDCS, which further confirms the successful overexpression of the limiting factors and the identification of TF binding events using subMNase-Seq. Asterisks mark the locations of annotated Gal80 binding sites. **C** – Read coverage in G1 for the 2950 high confidence sub-nucleosomal peaks 1kb upstream of TSS with a 2-fold difference in at least one time-point between the 2 strains except during G1. Peaks are sorted in descending order based on the mean value per region. **D** – Quantification of the heatmap in C as the average coverage around the peak location for the 2950 sub-nucleosomal peaks identified, confirming that there were no differences in G1.

In order to identify sub-nucleosomal peaks genome-wide, only fragments smaller than 100bp were considered for analysis. These fragments should correspond to most TF binding events and exclude confounding effects of nucleosome sized fragments. Then, the analytical approach from Gutiérrez et al. was followed to identify genuine peaks, by setting a minimum number of reads across all samples[160] (detailed analysis in Methods sections). Using this approach, 7493 sub-nucleosomal peaks were identified, of which 750 had a 2-fold difference between the strains in G1. As such, these 750 peaks were excluded from downstream analysis to rule out some replication-independent effects. In order to identify differential TF binding events that could explain the observed differences in gene expression, only peaks that were located 1kb upstream of a TSS and had a 2-fold difference in at least 1 time-point except during G1 were selected (2950 peaks - Fig. 5.2C and 5.2D – G1 time-point is shown). These represent approximately 61% of all peaks identified in promoter regions, which suggests a dramatic effect of RT changes on TF binding dynamics.

These analyses confirm that the approach used allows the identification of defined regions of the genome (approximately 100bp or less – 5.2C) where there is an accumulation of sub-nucleosomal sized fragments, which potentially correspond to TF binding footprints and that the differential peaks selected were not affected in G1 (Fig. 5.2C-D).

5.2 - Genome-wide advance in replication timing has a profound effect on transcription-factor binding dynamics

In order to better understand the effect of SSDDCS overexpression on TF binding during S-phase, the fraction of differential peaks which either increased or decreased (log-normalised fold-change SSDDCS / *sml1Δ* > 1 or < -1, respectively) in each time-point were identified (Fig. 5.3A). Overall, a higher percentage of peaks with decreased signal in SSDDCS was identified compared to peaks with increased signal (Fig. 5.3A). Surprisingly, 17% of the 2950 differential peaks showed decreased signal in early S-phase (5 minutes after release from G1). This could be explained by the fact that during replication fork progression, nucleosomes and TF

binding events have to be reassembled behind the fork. As such, higher rates of initiation in early S-phase upon SSDDCS overexpression could cause the ejection of TFs from their DNA binding locations and cause the observed differences.

In order to determine the effect of differential TF binding on gene expression, the proportion of differential peaks in the promoters of genes whose expression changed after a global advance in RT (Fig. 3.8 – k-means DE clusters) was analysed. This analysis showed that there was not a significant increase in the proportion of differential peaks in promoters of DE genes compared to the whole genome, except for the DE genes in cluster 3 (p-value = 0.03267, ~69% of peaks in promoter regions of genes from cluster 3 are differential peaks, compared to the 61% of peaks in promoters of all genes – Fig. 5.3B).

Analysis of the proportion of genes from each cluster that had differential sub-nucleosomal peaks in the promoter region provided similar results (Fig. 5.3C). 38% of all genes in the genome have at least one differential peak in the promoter region and none of the k-means had a significantly different result, except for cluster 3 (p-value = 0.0031, 47% of genes from cluster 3 have at least one differential peak in the promoter).

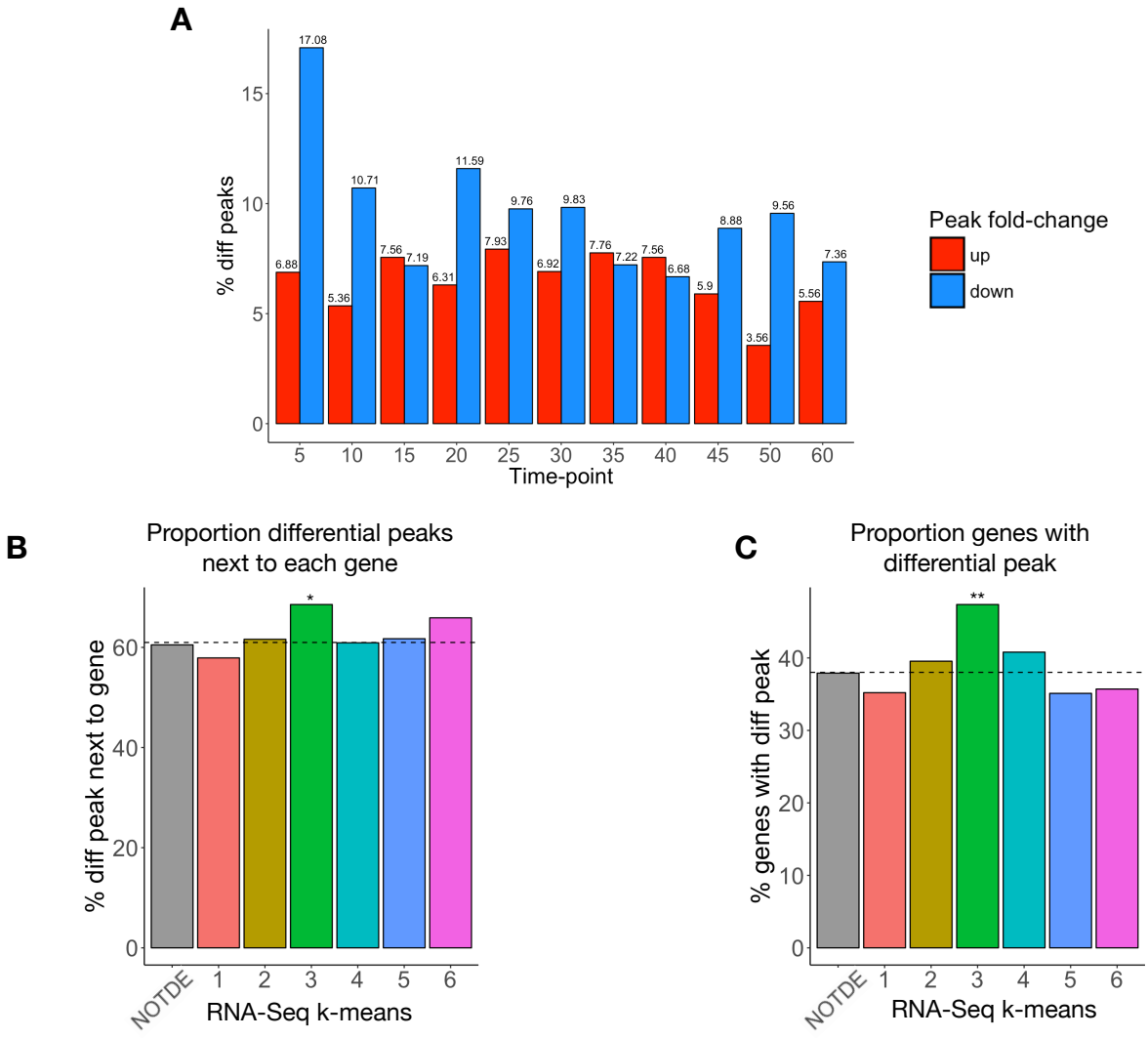


Figure 5.3 – Genome-wide advance in replication timing has a profound effect in TF binding dynamics, but all groups of DE genes show a similar effect compared to non-DE. A – Ratio of differential peaks which either increased or decreased (log_2 fold-change SSDDCS / *sml1Δ* > 1 or < -1 respectively) over all 2950 differential peaks located 1kb upstream of TSS in each time-point. Numbers on top of the bars represent the proportion of differential peaks in each time-point. **B** – Ratio of differential peaks over total number of peaks within promoter regions of genes from each DE k-means cluster. Dashed vertical line represents the ratio of differential peaks over all peaks identified in promoter regions of all genes (61%). **C** – Proportion of genes from each DE k-means cluster with differential peaks in the promoter region. Dashed vertical line represents the ratio of all genes in the genome that have at least one differential peak in the promoter region (38%). p-values are from an exact binomial test, ** $p < 0.01$, * $p < 0.05$.

Despite the statistically significant differences observed for cluster 3, these analyses did not show a particular effect on peaks regulating different classes of DE genes. The results presented so far suggest that the genome-wide advance in RT causes profound effects on gene expression, chromatin and TF binding dynamics, but that many of these events are not mechanistically linked, as similar proportions of non-DE and DE genes were identified with chromatin defects (Fig. 4.5A-B) and differential TF binding events (Fig. 5.3B-C).

One of the limitations of the unbiased subMNase-Seq approach is that it is not possible to determine which TF the identified peaks correspond to. MNase-Seq lacks the sensitivity and specificity of ChIP-Seq which allows the identification of regions of enrichment for a single TF. In order to verify that the sub-nucleosomal peaks identified potentially correspond to genuine TF binding events, peaks were annotated to TF binding sites from the map of regulatory sites generated in MacIsaac et al[180]. Only peaks within 200bp of an annotated site present in this map were considered for downstream analyses. In order to validate this approach, the GAL1-10 promoter was used, as peaks annotated to binding events in this promoter should have an increase in signal upon SSDDCS overexpression in all time-points. Moreover, these peaks should be among the strongest hits present in this dataset, due to the artificial overexpression of SSDDCS.

In order to address this possibility, the total number of reads mapping to each peak (as a sum of the reads overlapping each peak position across all samples) was plotted as a function of the $\log_2$ fold-change (SSDDCS / *sml1⊿*) in every time-point. As expected, 3 peaks were identified within 200bp of annotated Gal80 binding sites which showed an increased signal in the SSDDCS strain throughout the entire time-course and were among the peaks with higher read coverage (Fig. 5.4 – Gal80 peaks). These 3 peaks correspond to the 3 asterisks in figure 5.2B, which further validates the analytical approach used. This analysis also illustrates the greater number of peaks with decreased (blue) compared to increased (red) signal upon SSDDCS overexpression (Fig. 5.3A). A high threshold of total read coverage (sum of reads across all samples) was used in the volcano plots to make sure that the Gal80 peaks were identified (Fig. 5.4 - dashed horizontal line).

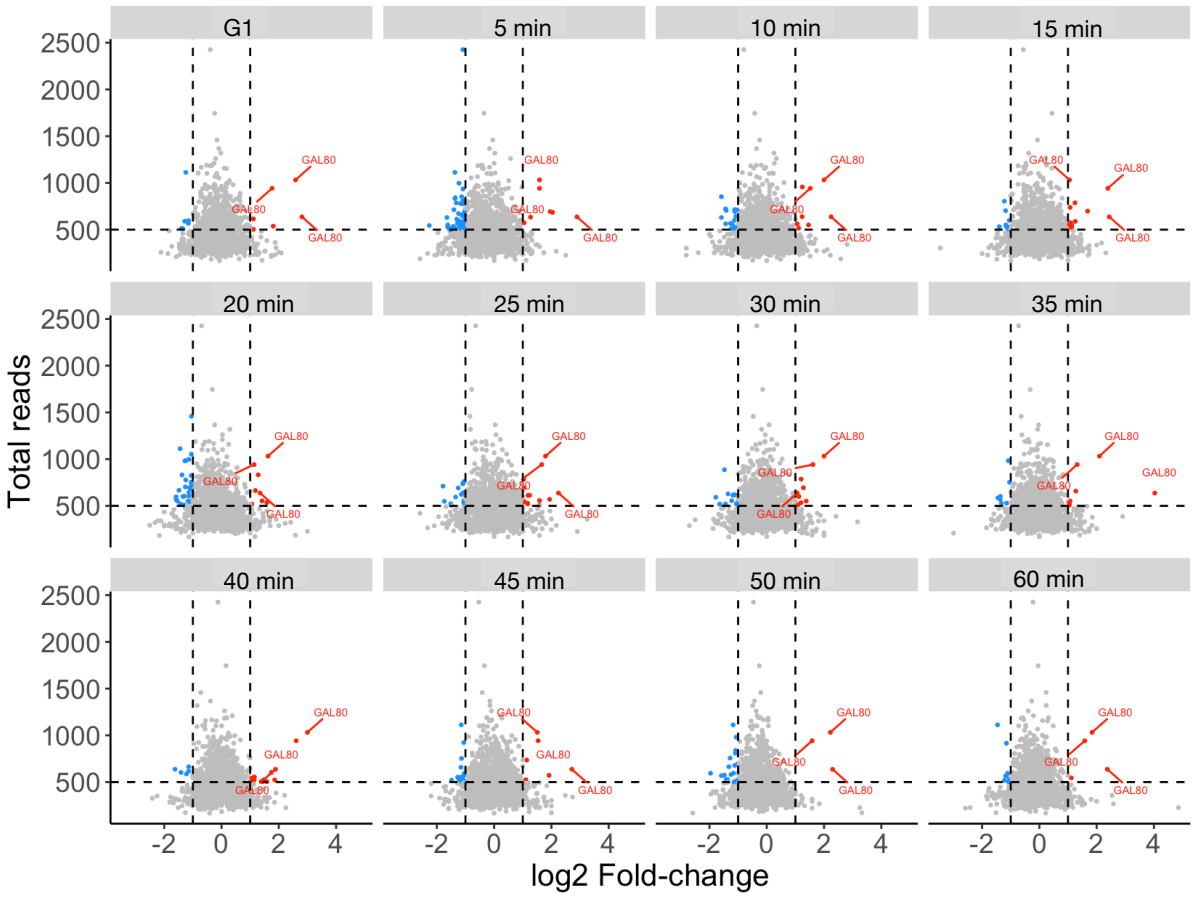


Figure 5.4 – Annotation of sub-nucleosomal peaks to known TF binding sites allows the identification of genuine TF binding events. Volcano plot of total number of reads as sum of all reads overlapping each peak position across all samples vs log_2 fold-change (SSDDCS / *sml1Δ*) in each time-point. Only peaks within 200bp of annotated TF sites from MacIsaac et al.[180] are shown. Gal80 peaks are highlighted to further validate the analytical and experimental approaches used. Dashed vertical lines marks the differential peaks with increased or decreased signal (log_2 fold-change SSDDCS / *sml1Δ* > 1 or < -1 respectively), and dashed horizontal line sets the threshold of sum of reads across all samples used = 500. This threshold was set for the purpose of identifying the differential peaks with higher read coverage, and as expected these include the Gal80 peaks.

It is important to mention that figure 5.4 has more peaks than the 2950 from figure 5.2C and 5.3, because differential peaks in G1 were excluded from the initial analyses. These include the Gal80 peaks, as the overexpression of SSDDCS is induced during the G1 arrest so the increased signal is observed in G1 (Fig. 5.4 – G1 time-point). The purpose of figure 5.4 is the validation of the annotation to known TF binding sites, which provided the expected results. The annotation of sub-nucleosomal peaks to known TF binding sites, combined with the identification

of high confidence peaks by setting a minimum read coverage across all samples, increases the chances that the peaks identified correspond to genuine TF binding events.

To address how well the data covered each TF, the number of sub-nucleosomal peaks annotated to each TF was compared to the number of sites identified in MacIsaac et al[180]. Strikingly, the number of peaks detected annotated to each TF was positively correlated with the number of sites identified in MacIsaac et al. (Fig. 5.5), suggesting that most TFs have similar coverage and that there was no bias towards specific TFs (i.e. TFs with more annotated sites have more sub-nucleosomal peaks identified and vice-versa). As such, this analysis shows that TFs with a low number of MNase-Seq peaks identified can still be considered for analysis, as this is a consequence of the small number of annotated binding sites genome-wide.

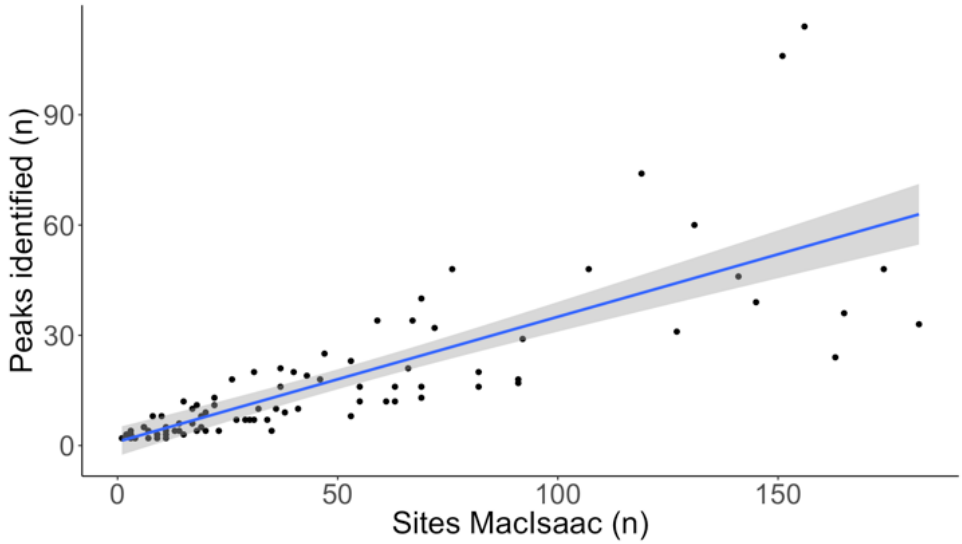


Figure 5.5 – Number of sub-nucleosomal peaks identified for each TF is directly proportional to the total number of annotated binding sites. Total number of peaks annotated to each TF vs number of annotated sites from MacIsaac et al.[180] Each data point corresponds to one TF. Blue line – linear regression line.

Among the 2950 differential peaks identified, 863 were located within 200bp of annotated TF binding sites, so these were the ones considered for downstream analysis. If the changes in peak intensity were a consequence of advanced replication, there could be a stronger effect in peaks located closer to origins of replication. However, plotting the $\log_2$ fold-change (SSDDCS / *sml1Δ*) as a function of peak distance to origins has shown that many peaks located very close to origins were not affected, and that differential peaks can be located within a wide range of distance from origins (Fig. 5.6).

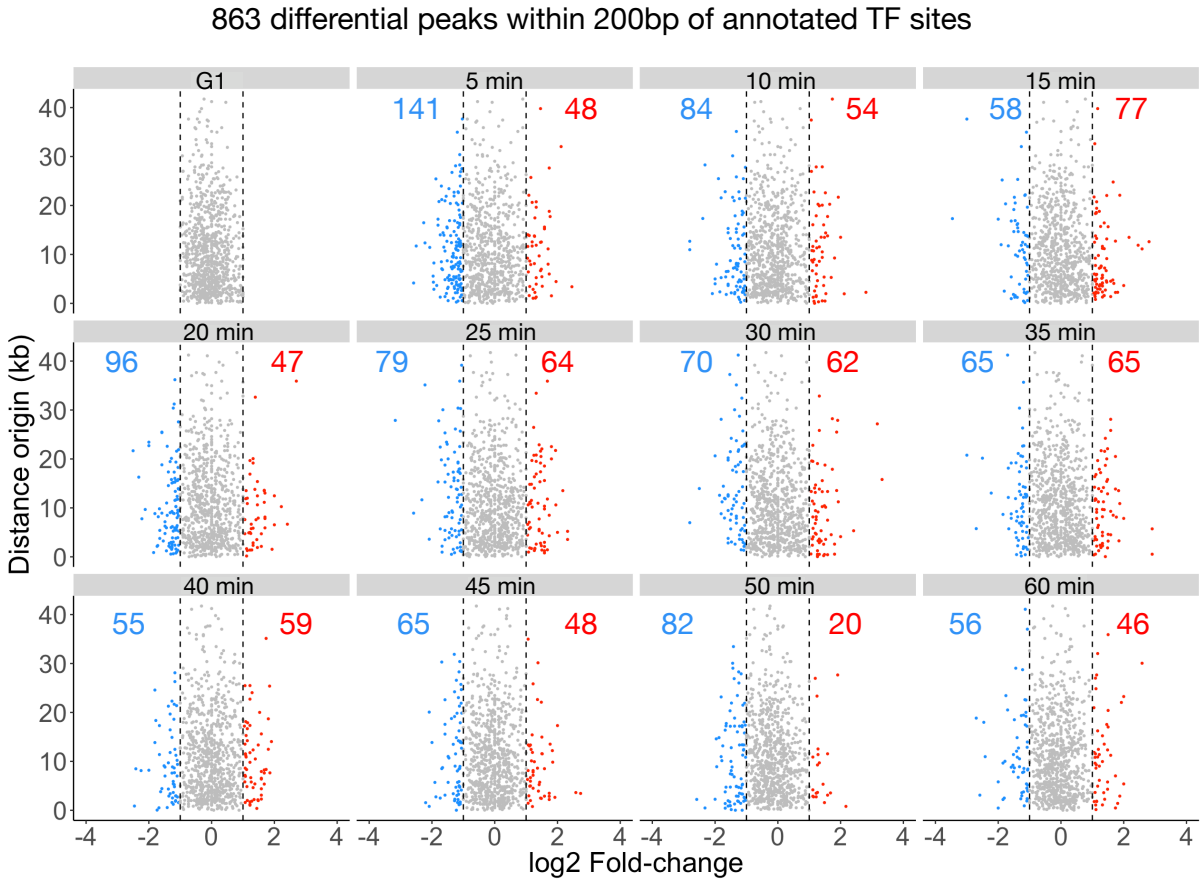


Figure 5.6 – Distance to origins is not a major determinant of the observed differences in sub-nucleosomal peak signal. Volcano plot of distance to closest origin vs $\log_2$ fold-change (SSDDCS / *sml1Δ*) in each time-point. From the 2950 differential peaks identified, 863 were within 200bp of annotated TF sites from MacIsaac et al.[180] The total number of differential peaks in each time-point is shown. As expected, there were no differential peaks in G1, as these were excluded from the analysis.

Comparison of the total number of differential peaks in each time-point indicates that overall there were more peaks with decreased signal in the SSDDCS compared to peaks with increased signal (Fig. 5.6), which is consistent with previous analysis (Fig. 5.3A). The results presented so far in this chapter validate the experimental and analytic approaches used for the identification of TF binding events and illustrate the genome-wide impact in TF binding dynamics upon global advance of RT.

5.3 - Different families of TFs are affected, including TFs involved in meiotic gene expression regulation

In order to identify changes in sub-nucleosomal peaks that could explain the observed differences in gene expression, the fraction of differential peaks annotated to different TFs which were located in the promoters of DE genes was analysed as follows:

1 - First, to identify TFs that have the greatest number of events affected upon the advance in RT, the ratio of differential peaks over all peaks annotated to each TF was calculated. 1538 peaks were identified in promoters regions of genes (1kb upstream of TSS) and within 200bp of annotated TF sites. Of these, 863 (56%) had either increased or decreased signal ("differential peaks", 2-fold difference between the two strains in one or more time-points except G1 – Fig. 5.6), so TFs that have a percentage of differential peaks greater than 56 are significantly more affected.

2 - Then, for each TF, the ratio of differential peaks in the promoter of DE genes over differential peaks in the promoter of non-DE genes was calculated.

These two ratios allow the simultaneous comparison of the effect of advancing RT on TF binding and gene expression. This analysis is summarised in Figure 5.7 and TFs with only one peak identified were excluded from this analysis. The dashed lines in figure 5.7 mark the ratios when considering all peaks (x-axis – 56% of all peaks annotated to TF sites and in promoters were affected in SSDDCS; y-axis – 35% of all TF differential peaks identified were in promoters of DE genes).

For example, the TFs with the highest ratio of differential peaks in promoter regions of DE genes were *SUM1*, *NDD1* and *ARG80*, as these have twice as many

differential peaks in promoters of DE genes compared to non-DE (Fig. 5.7 – y-axis = 2). On the other hand, *OPI1*, *SNF1*, *IME1*, *ADR1*, *SBT2* and *RFX1* were the TFs with the highest number of differential peaks compared to non-affected peaks (Fig. 5.7 – x-axis = 1), i.e., all peaks identified in this dataset annotated to these TFs were differential, so the ratio of differential over non-affected peaks was equal to 1.

The most interesting targets from this analysis would be the ones on the top right quadrant of figure 5.7: high number of differential peaks compared to total peaks (x-axis – high TF binding effect) and high number of differential peaks in the promoter of DE genes compared to non-DE (y-axis – high effect on DE genes). *RFX1* is a good example, as all peaks identified were differential (x-axis = 1) and there were 1.5 more differential peaks in the promoters of DE genes compared to non-DE (y-axis = 1.5). Only 5 peaks were mapped to *RFX1* (Fig. 5.7 – size of data points), which is expected considering that there are only 6 *RFX1* annotated sites in the MacIsaac dataset and that the number of sites identified was proportional to the number of annotated sites, as described previously (Fig. 5.5).

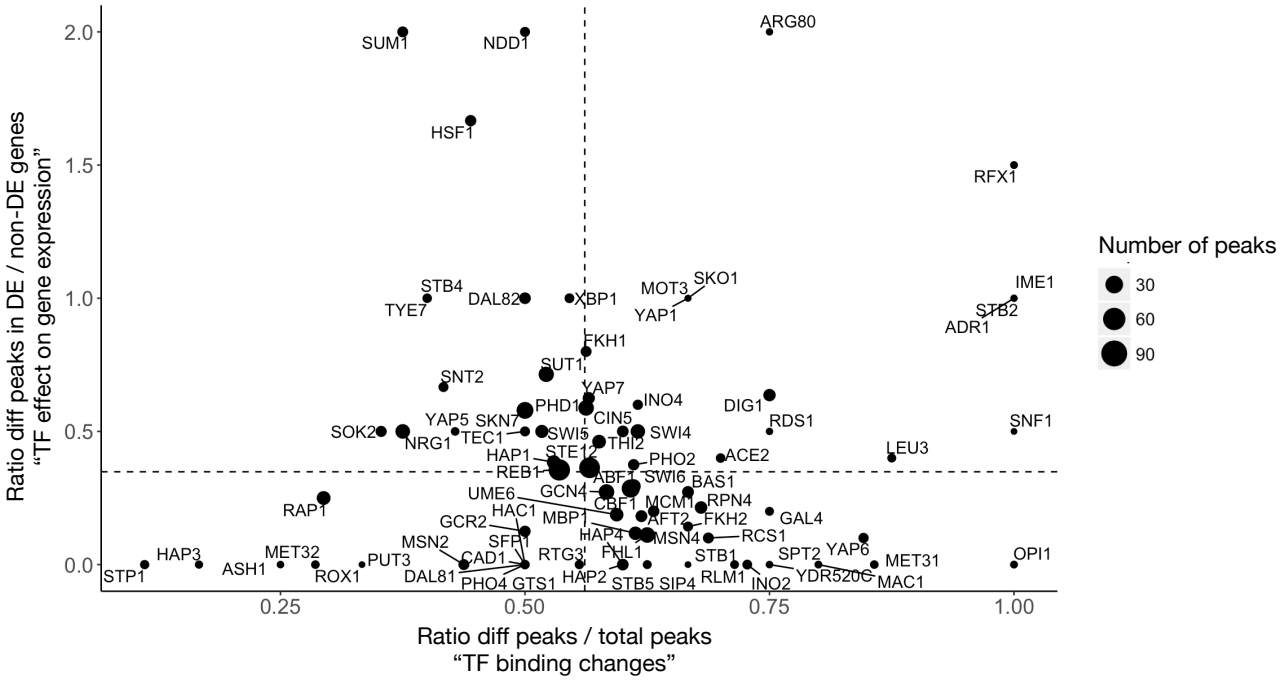


Figure 5.7 – Genome-wide analysis of impact of advancing RT on TF binding and consequent impact on gene expression. A - Ratio of differential peaks over total number of peaks (x-axis "TF binding changes") for each TF vs ratio of differential peaks in promoters of DE genes over differential peaks in promoters of non-DE genes (y-axis "TF effect on gene expression"). The size of the points is proportional to the total number of sub-nucleosomal peaks identified annotated to each TF. The dashed lines mark the ratios when considering all peaks. See main text for details.

This analysis allowed the identification of TFs with a significant number of differential binding events in the promoter of DE genes and illustrates the genome-wide impact of advancing RT on TF binding dynamics.

To gain further insight into the impact of differential TF binding dynamics on gene expression, individual groups of DE genes were analysed. RNA-Seq k-means cluster 1 and 3 were chosen for the following reasons: cluster 1 genes have been used during this work to dissect the relationship between RT, transcription and chromatin, for reasons already described in previous chapters, while cluster 3 was chosen because it was the only cluster with a significant enrichment of genes with differential sub-nucleosomal events (Fig. 5.3 B-C). As such, the ratio of differential peaks over total number of peaks (TF binding effect) determined previously was compared with the ratio of differential peaks in promoters of genes from cluster 1 and 3 over differential peaks in promoters of any DE gene (Fig. 5.8)

Once again, TFs in the top right quadrant represent the ones with higher number of differential binding events in promoters of a higher proportion of genes from the same cluster. For *RFX1*, as an example, all the peaks annotated to this TF were differential (x-axis = 1) and were all in the promoter of genes from k-means cluster 1 (y-axis = 1) (Fig. 5.8 – top). Rfx1 is a major repressor of DNA damage regulated genes that recruits the Tup1/Cyc8 repressor complex already described[181] (Fig. 4.9A).

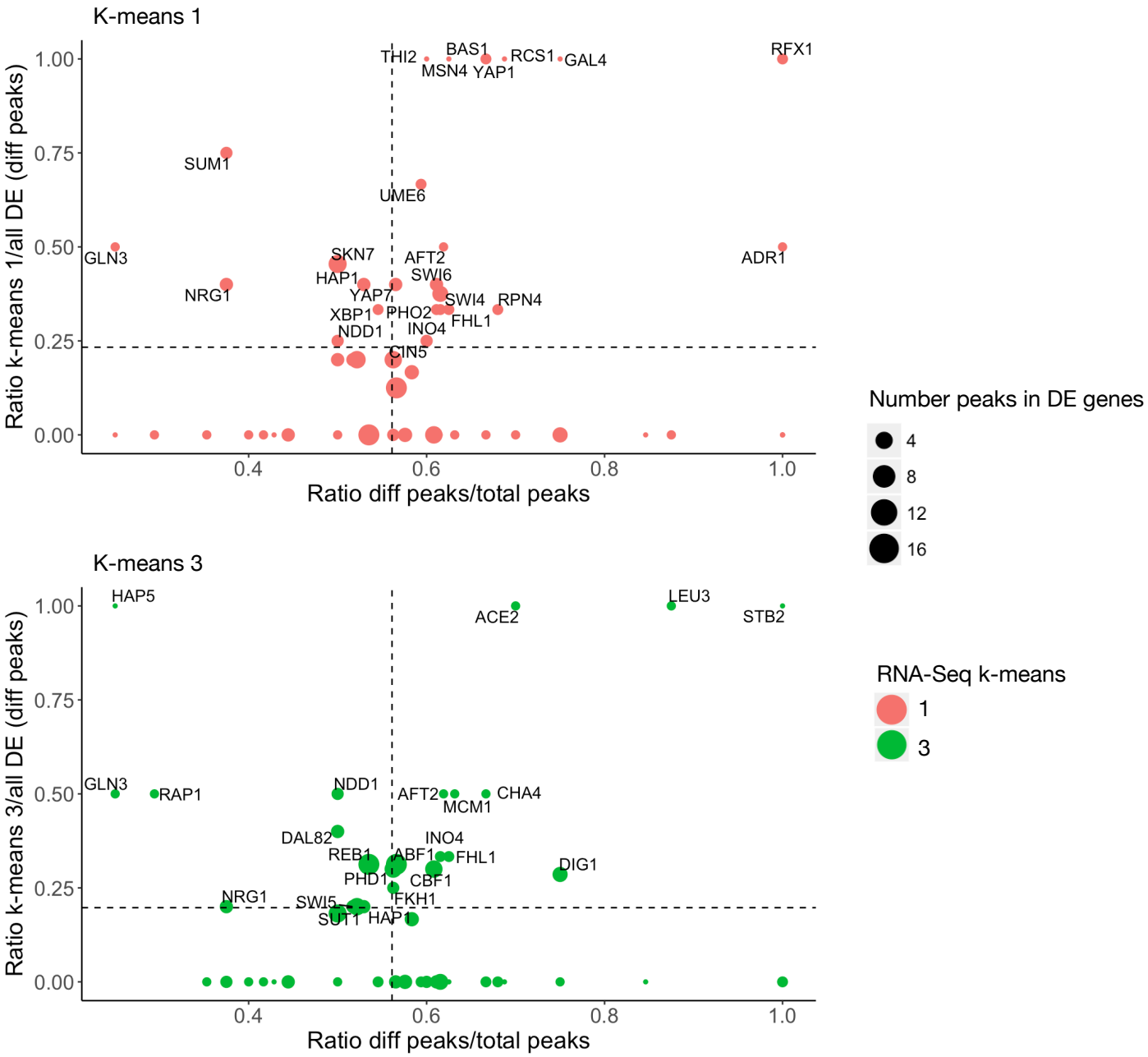


Figure 5.8 – Analysis of impact of advancing RT on TF binding in the promoter of genes from k-means cluster 1 and 3. - Ratio of differential peaks over total number of peaks (x-axis) for each TF compared to ratio of differential peaks in promoters of DE genes from k-means clusters 1 and 3 (top and bottom respectively, plots were coloured using the corresponding k-means colours as done previously) over differential peaks in promoters of any DE gene (y-axis). The size of the data points is proportional to the total number of peaks from each TF in promoters of DE genes to help with the interpretation. The vertical dashed line marks the proportion of all differential peaks over all peaks (56%). The dashed horizontal line marks the proportion of differential peaks in promoters of genes from each k-means cluster over differential peaks in promoters of all DE genes. Only TFs with a y-axis ratio higher than the whole genome proportion are labelled (23% and 20% of TF differential peaks in promoters of DE genes are in promoters of genes from cluster 1 and 3, respectively).

Another example that illustrates the power of this analysis to identify differential peaks from the same TF affecting different genes is *ADR1*: all peaks were differential, but only 50% of the peaks in the promoters of DE genes were in promoters of genes from cluster 1 (Fig. 5.8 top – y-axis = 0.5), while the other 50% were in promoters of genes from cluster 5 (not shown). As such, the ratio calculated in the y-axis provides an indication of the proportion of differential events from each TF regulating the different classes of DE genes.

As described in previous chapters, SSDDCS overexpression induces the activation of meiotic genes (Fig. 3.8 and 3.11) together with a perturbation of chromatin around Ume6 binding sites, which is involved in the silencing of these genes (Fig. 4.9 and 4.10), suggesting an impact in the binding of Ume6. A closer look at the regulators of the meiotic genes, showed that despite the fact that only 3 out of the 19 *UME6* differential peaks were in promoters of DE genes (19%, y-axis = 0.19 Fig. 5.7), 2 of these were in promoters of genes from cluster 1 (67%, y-axis = 0.67 Fig. 5.8 top) while the other one was in the promoter of a gene from cluster 5 (not shown). Another example of a negative regulator of meiotic genes is *SUM1*[182], and 3 out of the 4 differential peaks in promoter of DE genes were in genes from cluster 1 (75%, y-axis = 0.75 Fig. 5.8 top), which is in agreement with the functional enrichment of meiotic genes in this cluster.

Notably, *STB2* was among cluster 3 top hits (Fig. 5.8 bottom): Stb2 is a DNA binding protein that interacts with Sin3[183], which is part of the Rpd3 histone deacetylase complex already described, that acts as a transcriptional repressor of several processes, including meiosis[174] (Fig. 4.9B). However, only 2 *STB2* peaks were identified and despite the fact that they were both differential (hence x-axis = 1, Fig. 5.8 bottom), only one was in the promoter of a DE gene and this gene was not involved in meiosis regulation (and this gene belonged to cluster 3, hence y-axis = 1 Fig. 5.8 bottom). Still, these results suggest that the binding patterns of TFs that are part of the same DNA binding complexes were affected, illustrating the complexity of the impact of advancing RT on the TF binding landscape.

Another top hit among cluster 3 genes was *ACE2,* a TF involved in septum destruction after cytokinesis: *ACE2* mRNA expression is highly periodic and peaks

in G2/M. Moreover, Ace2 is actively exported from the nucleus during cell cycle phases other than cytokinesis[184], making it difficult to explain the differential events observed during S-phase. Still, its identification suggests that advancing RT could have an impact on the homeostasis of cell cycle regulated TFs.

Finally, the analyses presented so far considered differential peaks as a whole without distinguishing whether the differential peaks had an increase or decrease in signal upon SSDDCS overexpression and whether the effect was present during a single time-point or several. Considering that genes belonging to each k-means cluster have similar expression profiles (Fig. 3.8), it is possible that sub-nucleosomal peaks annotated to the same TF and regulating genes from the same cluster have similar binding dynamics.

Analysis of the fold-change per time-point (Fig 5.9) allowed the identification of some of these events, but similar binding patterns for peaks annotated to particular TFs were not observed overall (Fig. 5.9). Also, many TFs had only one differential peak per k-means cluster and many differential peaks were affected in late time-points, which could still be caused by the RT advance but is unlikely to have an impact on expression of the DE genes identified, as most changes in transcription took place during early to mid S-phase (Fig. 3.7 and 3.8). Taking *RFX1* and DE genes in k-means 1 again as an example, two differential peaks showed decreased binding early in S-phase, while another showed increased binding in late S-phase (Fig. 5.9 left). The *SUM1* differential peaks in cluster 1 genes also presented different dynamics, as all peaks showed a decreased signal but in different periods of S-phase (one early, one middle and one in late S-phase – Fig 5.9 left). Finally, the two *UME6* peaks regulating genes from cluster 1 had opposite binding patterns in early S-phase. The overall lack of similar binding patterns was also observed for the differential sub-nucleosomal peaks regulating genes from k-means cluster 3 (Fig. 5.9 right), despite the two *ACE2* peaks with decreased signal during mid S-phase (Fig. 5.9 right).

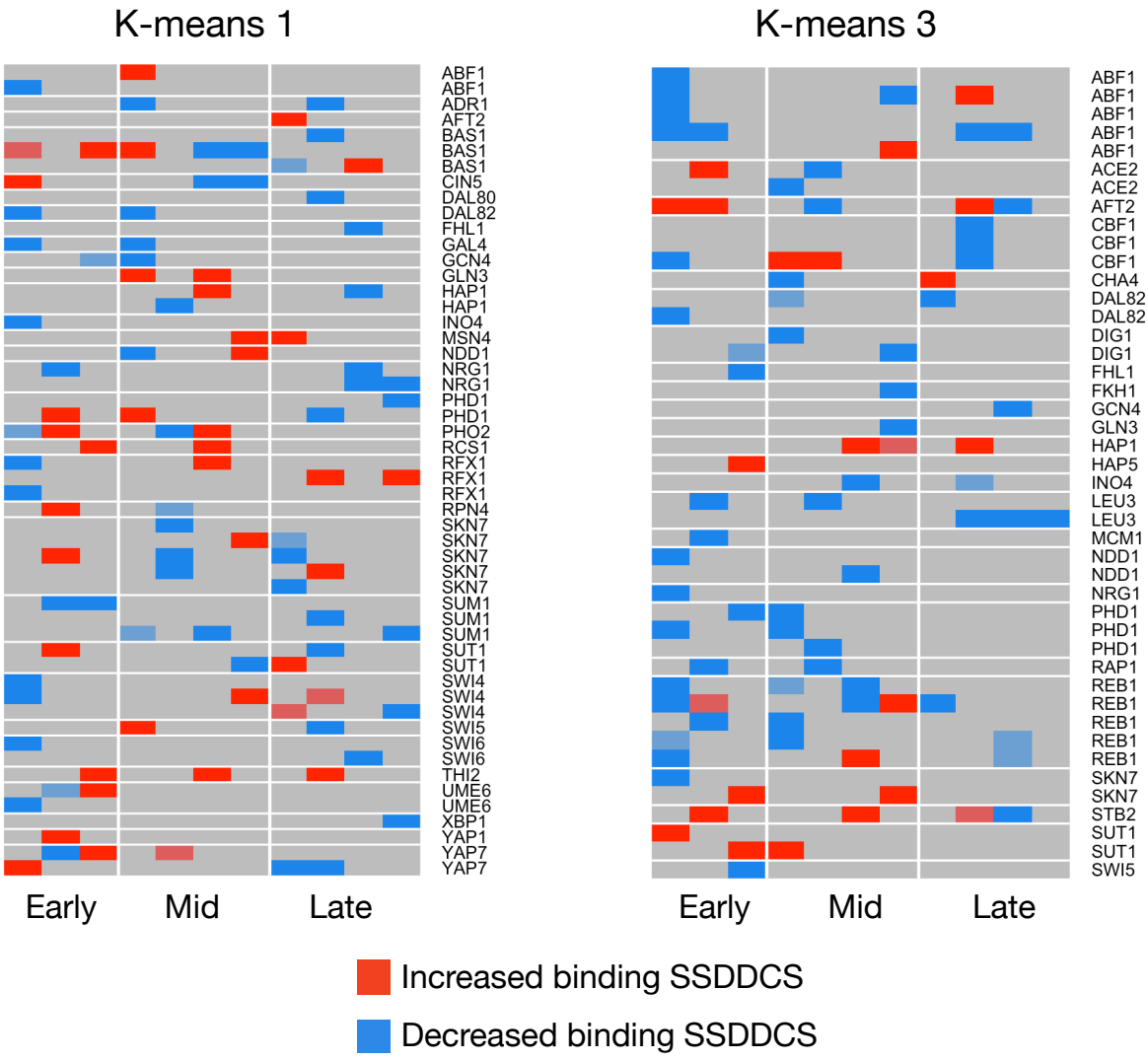


Figure 5.9 – TF binding dynamics during S-phase upon RT advance. – Peak $\log_2$ fold-change (SSDDCS / *sml1Δ*) for the differential sub-nucleosomal peaks annotated to known TF sites regulating genes from k-means cluster 1 and 3 (as in Fig. 5.8). Each row represents one peak and each column one time-point, from 5 to 60 minutes after G1 release (as done previously). To help visualisation and interpretation, time-points were split into three S-phase periods (Early – 5, 10, 15; Mid – 20, 25, 30, 35 and Late – 40, 45, 50, 60) and divided using white vertical lines. Peaks were coloured in red or blue depending if the $\log_2$ fold-change in that time-point was higher or lower than 1, respectively. If the peak was not affected in a particular time-point, it was coloured in grey. The TFs annotated to each peak are displayed on the right side of the heatmaps. Peaks were sorted alphabetically by TF, so that different TFs can be directly compared, and white horizontal lines delimit peaks from the same TF.

The lack of consistent trends observed could be a consequence of the already described limitations of the approach or simply because differences in TF binding were not the major determinant of gene expression changes in the experimental system used. Another caveat of the approach of comparing TF binding and gene expression is the fact that the changes in TF binding are expected to precede the changes in mRNA, because of the delay between TF binding and initiation of transcription. Moreover, many of these TFs act as scaffolds that recruit several activators and repressors of gene expression, and it would not be possible to distinguish different TFs binding in the same genomic region using this approach.

Finally, the candidate genes analysed in detail in previous chapters were also used for the identification of differential TF binding events that could explain the observed differences in gene expression and chromatin organisation. No high-confidence sub-nucleosomal peaks were identified in the promoters of *IME2* or *RIM15*, but one of the Ume6 peaks targeting genes from cluster 1 was in the promoter of *NDT80*, and this peak showed decreased signal in the SSDDCS strain 5 minutes after G1 release (Fig. 5.10A).

As described previously, Ume6 acts as a repressor of meiotic genes during mitosis, and considering that a differential TF binding event would precede changes in transcription, a decrease in Ume6 binding early in S-phase could explain the up-regulation of *NDT80* in the SSDDCS strain in later time-points (20-30 minutes after G1 release - Fig. 5.10B). Moreover, a movement of nucleosomes close to the Ume6 binding sites in *NDT80* promoter was also observed upon SSDDCS (Fig. 5.10C) and the origin deletion experiments have shown that the up-regulation of *NDT80* is most likely a direct consequence of the advance in replication (Fig. 3.13B and 3.13E). As such, this locus showed a clear association between RT, gene expression, chromatin organisation and TF binding (Fig. 5.10).

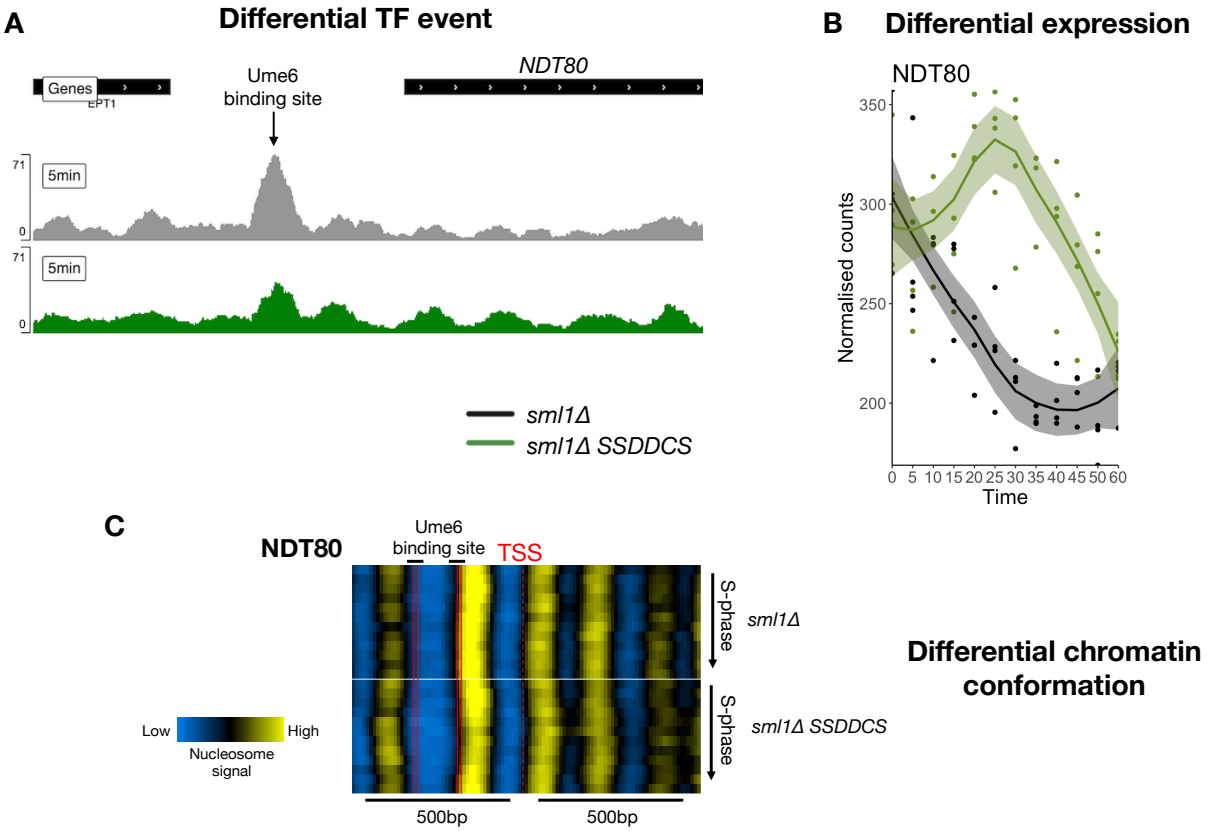


Figure 5.10 – The overexpression of SSDDCS impacts transcription, chromatin and TF binding landscape of the *NDT80* locus. A – IGV track of *NDT80* locus at 5 minutes after G1 release. Y-axis shows the sub-nucleosomal read coverage in this genomic region. Top track shows the location of genes. Peak annotated to Ume6 binding site is highlighted. **B** – Gene expression profile of *NDT80*, as shown in chapter 3 (Fig. 3.12B). **C** – Nucleosome profile centred around *NDT80* TSS and highlighting Ume6 binding sites, as shown in chapter 4 (Fig. 4.10).

Some other meiotic genes also showed differential binding events that could explain the differences in expression observed: Sum1 acts as a repressor of meiotic genes during mitosis as described previously. A Sum1 peak with decreased signal in the SSDDCS strain was identified 15 minutes after G1 release in the promoter of *YSW1*, a gene required for normal spore membrane formation[185] (Fig. 5.11A). The decreased binding of the Sum1 repressor could also explain the up-regulation of this gene upon SSDDCS overexpression (Fig. 5.11B). However, no changes in chromatin organisation were observed in *YSW1* locus (not shown).

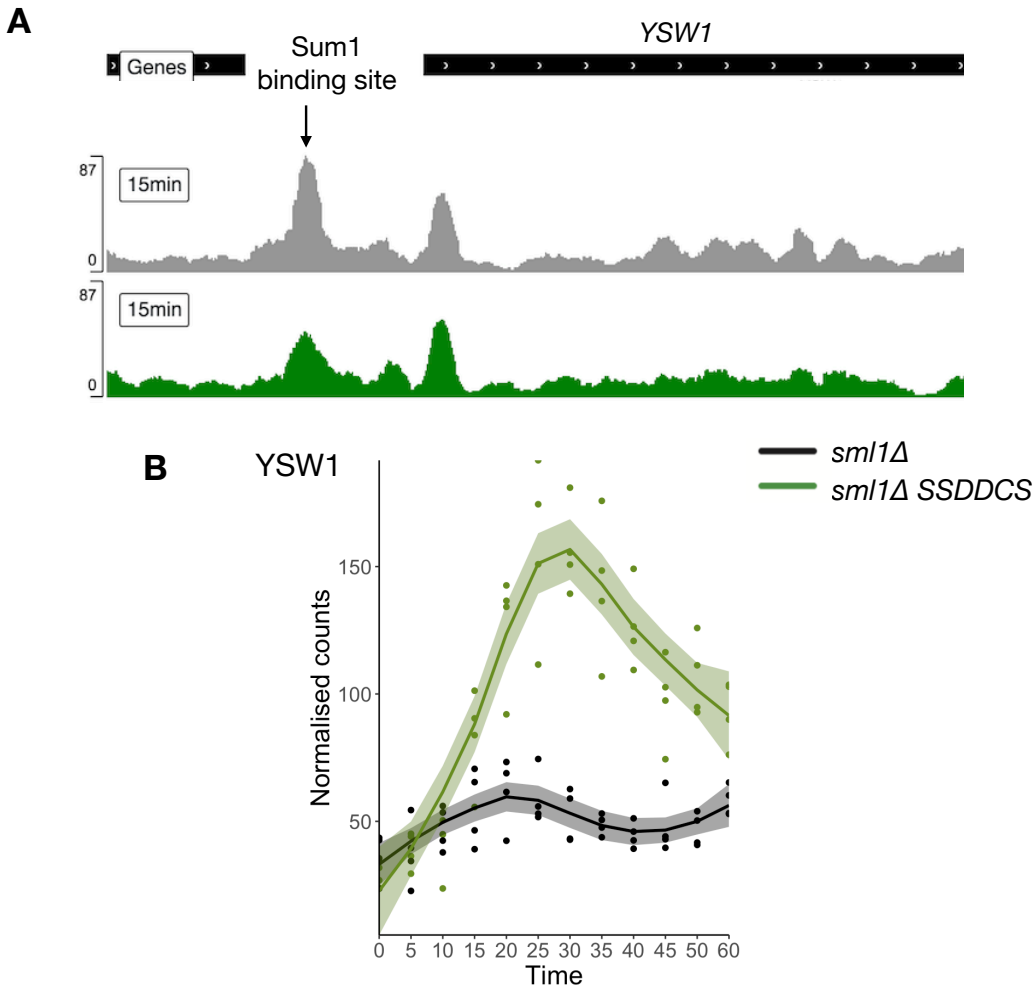


Figure 5.11 – The overexpression of SSDDCS impacts transcription and TF binding landscape of *YSW1* locus. A – IGV track of *YSW1* locus 15 minutes after G1 release. Y-axis shows the sub-nucleosomal read coverage in this genomic region. Top track shows the location of genes. Peak annotated to Sum1 binding site is highlighted. **B** – Gene expression profile of *YSW1*.

Finally, there were also cases of sub-nucleosomal events with increased signal in the SSDDCS strain (rather than decreased) in the promoter of meiotic genes, such as *LDS1,* a meiotic gene involved in spore wall assembly[186] (5.12A). However, this binding event was not annotated to any known meiotic regulator but to *SWI5*, a TF involved in regulation of genes expressed during M/G1 which is a *ACE2* paralog[184]. This could be explained by several reasons, such as errors during the annotation of binding sites or different TFs binding in the same region. Still, if this event corresponds to the activator of *LDS1*, this differential binding event could be the cause of the increase in expression (5.12B). Moreover, *LDS1* showed nucleosome movement in the gene body coincident with the increase in transcription and TF signal (Fig. 5.12C).

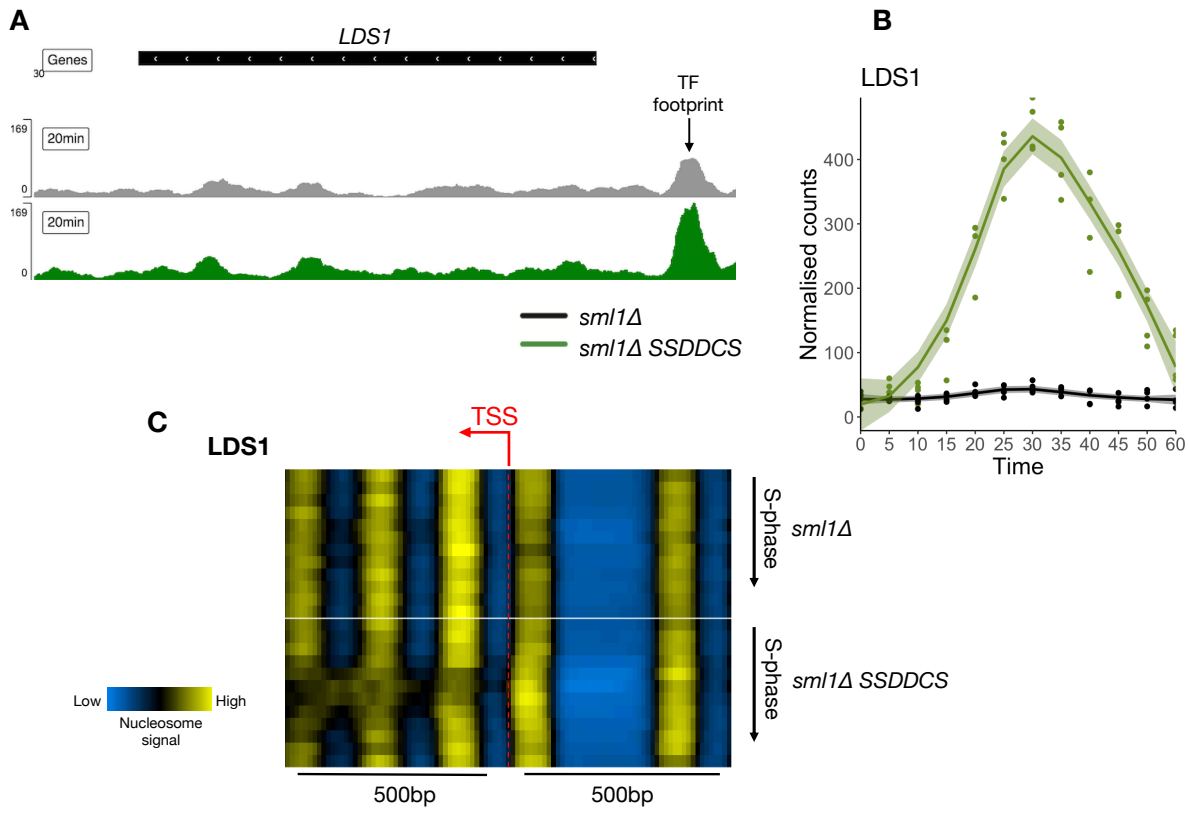


Figure 5.12 – The overexpression of SSDDCS impacts transcription, chromatin and TF binding landscape of *LDS1* locus. A – IGV track of *LDS1* locus 20 minutes after G1 release. Y-axis shows the sub-nucleosomal read coverage in this genomic region. Top track shows the location of genes. The differential TF footprint (annotated as *SWI5* in MacIsaac et al.) is highlighted. **B** – Gene expression profile of *LDS1*. **C** - Nucleosome profile centred around *LDS1* TSS. Notice that *LDS1* is located in the – DNA strand, so the direction of transcription is from right to left. There is movement of the nucleosomes in the gene body (left side of heatmap) upon SSDDCS overexpression.

Despite the overall lack of association between differential expression of genes and differential TF binding events (Fig. 5.8 and 5.9), there were cases of single genes where perturbations in TF binding were associated with changes in chromatin conformation and gene expression (Fig. 5.10 - 5.12). Moreover, the origin deletion experiments presented in this thesis have shown that the up-regulation of *NDT80* is most likely a direct consequence of the advance in RT of this locus (Fig. 3.13B and 3.13E), which might impact the chromatin conformation in the binding sites of a known regulator (Ume6, Fig. 5.10C) and consequently, the correct binding to *NDT80* promoter (Fig. 5.10A).

The results presented during this chapter illustrate the profound dysregulation of the TF binding landscape throughout S-phase upon global advance in RT, with several classes of TFs affected (Fig. 5.6 – 5.9). This is consistent with the effects observed on gene expression and chromatin.

The subMNase-Seq results presented in this chapter would have to be validated with ChIP-Seq experiments of individual TFs, in order to confirm the changes observed. In the final chapter, I will discuss the limitations of our experimental approach and techniques used, as well as the most important results and potential mechanisms responsible for the observed changes. Overall, our results reinforce the importance of a defined order of origin firing to maintain the correct gene expression, chromatin organisation and TF binding patterns, and raise the possibility that regulation of replication timing could be used as a trigger to induce different cellular states.

Chapter 6 – Discussion

6.1 - Replication timing and genome homeostasis

DNA replication timing is associated with gene expression, and the link between the two was described more than 60 years ago. Therefore, it is surprising that despite the number of studies trying to dissect the details of this relationship, a complete mechanistic view of how RT affects transcription (and vice-versa) is still missing. This close association dictates that perturbations of either RT or transcription will have an impact on the other, turning this problem into a “chicken or the egg?” scenario. Moreover, most changes in RT and transcription during cellular differentiation take place almost in parallel or in a short temporal window, making it difficult to evaluate which one precedes the other. Despite the several advances in the field (described in the Introduction), experimental systems that allow controlled manipulations of RT with the temporal resolution required to dissect this close relationship are still missing.

In this thesis, I used a conditional system where 6 limiting initiation factors are overexpressed in budding yeast to advance RT genome-wide in a single cell cycle (Fig. 3.2). Overexpression of these factors is induced in a cell population arrested in G1 prior to release into a synchronous S-phase (Fig. 3.1). This temporal resolution together with the use of whole-genome NGS techniques allow a genome-wide view of replication dynamics and chromatin/transcription regulation. Moreover, budding yeast is one of the few organisms where origins of replication are defined by specific sequences, allowing further RT manipulations in defined regions of the genome. As such, this system was used to address the impact of a global RT advance on gene expression patterns (chapter 3), the chromatin landscape (chapter 4) and TF binding dynamics (chapter 5) during S-phase.

Perhaps not surprisingly, the global RT advance was associated with a dramatic impact on transcription, chromatin and TF binding, as follows:

1 – Expression of approximately 27% of all genes was affected and different clusters of DE genes with different expression profiles were identified, illustrating the heterogeneous impact of RT advance on transcription (Fig. 3.8).

2 – Chromatin organisation in gene bodies (Fig. 4.8) and movement of the +1 nucleosome (Fig. 4.6) were affected genome-wide, with more disorganised chromatin and more mobile +1 nucleosomes upon the RT advance.

3 – Binding events annotated to different families of TFs were affected, regulating different classes of genes (Fig. 5.7 and 5.8).

Strikingly, most of the observed changes were resolved once replication was complete: gene expression patterns (Table 3.1 and Fig. 3.7) and chromatin organisation in promoters and gene bodies (Fig. 4.6 and 4.8) were re-established by the end of S-phase. Moreover, there was less differential TF binding events in late S-phase, compared to earlier time-points (Fig. 5.6 and 5.9). These results suggest that high rates of origin firing early in S-phase have a transient impact that is resolved as initiation rates decrease and replication finishes.

6.2 – Replication timing and chromatin assembly

As described in the Introduction, higher rates of initiation early in S-phase cause the exhaustion of several factors such as the dNTP pool and topoisomerases. As such, the shorter S-phase induced by the SSDDCS overexpression “forces” cellular processes that take place during the full extent of S-phase to be completed in a much shorter window of time, such as the re-assembling of nucleosomes into chromatin once replication is complete. Therefore, the chromatin disorganisation observed upon SSDDCS overexpression could be caused by defects of histone deposition and assembly. Consistent with this hypothesis, unpublished work from Lukas Fiedler, a Part II student in the Zegerman lab which compared the gene expression profile of the SSDDCS strain with a chromatin machinery deletion compendium of 165 proteins[187], has shown that the transcriptomes of mutants of

histone chaperone complexes, such as members of the Chromatin Assembly Complex 1 (CAF-1), were similar to the transcriptome of the SSDDCS strain. There are some limitations to this comparison, such as the fact that the transcriptomes of these mutants were analysed in asynchronous populations using microarrays, compared to the synchronous S-phase / RNA-Seq approach used in this thesis. Moreover, the total number of DE genes identified was different between the datasets, despite the statistical significant associations: for example, the *cac1Δ* mutant had 77 DE genes (compared to >1700 in the SSDDCS strain) of which 46 were DE in the SSDDCS strain (Cac1 is part of the CAF-1 complex). Still, this analysis suggests that some of the changes observed in the SSDDCS strain could be due to defects on nucleosome assembly caused by the high rates of initiation in early S-phase, which would also explain the observed impact on chromatin organisation.

A decrease in the number of histones, which is observed during replicative aging, also causes profound changes in gene expression: Hu et al. observed that nucleosome occupancy decreased by 50% across the whole genome during replicative aging, together with the up-regulation of all budding yeast genes[188]. We did not observe a significant reduction on the total number of nucleosomes using our approach (Fig. 4.3), but unlike Hu et al., we did not use a spike-in to normalise our MNase-Seq results. To rule out that our SSDDCS strain causes a global reduction in nucleosome occupancy, we should repeat the MNase-Seq experiments with a spike-in control. Still, the MNase-Seq experiments present in this thesis were replicated 3 times and the DANPOS analysis pipeline used includes normalisation steps that allow the identification of differences in occupancy[159], and overall we have not identified such significant differences.

Moreover, Hu et al. compared “young” cells with “old” cells that have completed a median of 25 cell divisions, while we analysed the effects in a single cell cycle and identified DE genes with different expression profiles (both up and down-regulated). Still, differences in some of the up-regulated genes identified (such as k-means cluster 1) could be caused by a transient decrease in nucleosome occupancy due to defects on chromatin assembly once a locus is replicated. Hu et al. also showed that overexpression of histones H3/H4 partially reversed the changes in gene

expression[188], so it would be interesting to test whether overexpression of histones in combination with SSDDCS overexpression could revert some of the changes observed.

As described in the Introduction, the expression of histone genes increases during S-phase so that histone supply is sufficient for chromatin maturation. The increase in replication rates early in S-phase could cause a reduction in the level of histones relative to replicated DNA, which would have an impact on chromatin assembly. This could explain the higher disorganisation of chromatin observed in k-means cluster 1 genes (Fig. 4.8), as this is the group of genes located closest to origins and consequently the genes with the greatest RT advance (Fig. 3.10A).

Differences in chromatin maturation kinetics could also potentially explain the differences in gene expression observed if re-establishment of chromatin in DE genes is slower compared to non-DE, for example. Gutiérrez et al. used MNase-Seq to study chromatin maturation kinetics behind the replication fork and have identified regions with different maturation speeds[160], i.e. in some regions nucleosomes and TFs are immediately deposited behind the DNA replication fork, while other experience transient depletion of nucleosomes and TFs. However, Lukas Fiedler, the Part II student that compared the SSDDCS datasets with published datasets, found that the distribution of DE and non-DE genes according to maturation kinetics was mostly uniform, which suggests that the changes in gene expression are not dependent on different maturation kinetics (Zegerman lab unpublished observations).

An ideal system to study the impact of RT while maintaining the rates of origin firing constant during S-phase would be one where early origins fire late and late origins fire early, but such system is currently missing. One system that resembles this scenario the most is the RIF1 knock outs in human cell lines used in Klein et al.[99], which showed that the RT changes caused by loss of RIF1 were associated with alterations in chromatin modifications and the genome 3D structure (described in the Introduction). RIF1 is bound to late replicating domains and keeps these domains in the nuclear periphery, so the authors suggest that RIF1 deletion affects the number of origins competing for initiation factors. Consistent with this

hypothesis, the authors showed that loss of RIF1 affects RT by increasing RT heterogeneity between individual cells. The absolute impact on RT varied depending on the cell line used, but overall the effects included changes in replication domains from early to late and late to early and loss of defined initiation zones, causing approximately 40% of the genome to change RT[99]. Still, many regions retained the wild-type RT, suggesting the existence of RIF1-independent mechanisms regulating RT, even in late replicating domains. As described in the Introduction, RIF1 controls RT of telomeric regions in budding yeast, but a high resolution whole-genome replication profile of a *rif1Δ* yeast strain is still missing. As described previously, the metazoan genome is organised in replication domains spanning several kilo-bases, making local modulations of RT timing more challenging compared to the relatively simple origin deletion/insertion strategy in budding yeast. As such, a combination of RIF1 knock out / knock down with origin deletion/insertion in budding yeast could significantly perturb RT without increasing the simultaneous rate of origin firing during S-phase, allowing further insights into the role of RT on different genomic features such as transcription and chromatin.

6.3 - Replication timing and telomeric silencing

The significant advance in RT of sub-telomeric regions (Fig. 3.4) could also lead to defects in heterochromatin silencing due to problems in chromatin assembly, leading to dysregulation of genes in these regions. CAF-1 and Rtt106, which are among the top hits from the analysis comparing the SSDDCS transcriptome with the chromatin machinery deletion compendium described above, are involved in the formation of telomeric heterochromatin[189] and some clusters of DE genes were shown to be associated with telomeres, such as cluster 1 (Fig. 3.9B). Moreover, as described previously, cluster 1 showed the greatest effect on +1 nucleosome movement and chromatin organisation (Fig. 4.7C and 4.8E). The earlier replication of telomeres, which in the SSDDCS strain took place at the same time as the early replicating centromeres, could perturb chromatin deposition in these regions causing the observed up-regulation in gene expression.

6.4 - Replication timing and the TF binding landscape

As described in chapter 3, delaying the replication of the *NDT80* and *IME2* loci in an otherwise advanced S-phase (SSDDCS background) was sufficient to restore the wild-type expression levels of these two genes (Fig. 3.13). Moreover, chromatin was affected in the promoters of these two genes near to Ume6 binding sites (Fig. 4.10), the repressor protein responsible for controlling their expression. This result suggests that earlier replication of these loci could eject Ume6 from its binding sites, which might create a window of opportunity for these genes to be expressed. As such, there is the possibility that the delay of the RT of these regions allowed the timely reassembling of chromatin and establishment of the TF binding landscape, maintaining these genes repressed.

There are at least two other possible models where RT could affect TF binding patterns. The advance in RT caused by SSDDCS overexpression, which has a genome-wide effect on chromatin organisation, could out-titrate TFs from their native binding sites to sites that become transiently more accessible or cryptic sites which are normally not used (Fig. 6.1 - top). This could explain the differences observed in *IME2* and *NDT80* loci, as out-titration of Ume6 to sites other than the ones regulating these genes would lead to the up-regulation of these genes.

On the other hand, gene expression is highly cell cycle regulated in budding yeast[190] (a feature conserved in eukaryotes), in order to make sure that genes required for specific cell cycle phases are timely expressed. As such, advancing RT genome-wide could perturb the expression of genes which are targets of cell cycle regulated TFs. For example, a TF might have to accumulate during S-phase to reach the high levels required to activate the expression of its target genes in late S-phase. Moreover, if the target genes are late replicated, this could dictate the S-phase window where the promoters are accessible for binding. Advancing RT would make these promoters accessible early in S-phase, when the TF has not accumulated to the threshold required for proper regulation of the target genes, so the expression of these genes would be impacted (Fig. 6.1 – bottom).

It is important to mention that these models would have to be further tested and that the subMNase-Seq dataset provided some examples where the first model could be true but there was not enough evidence supporting the second model.

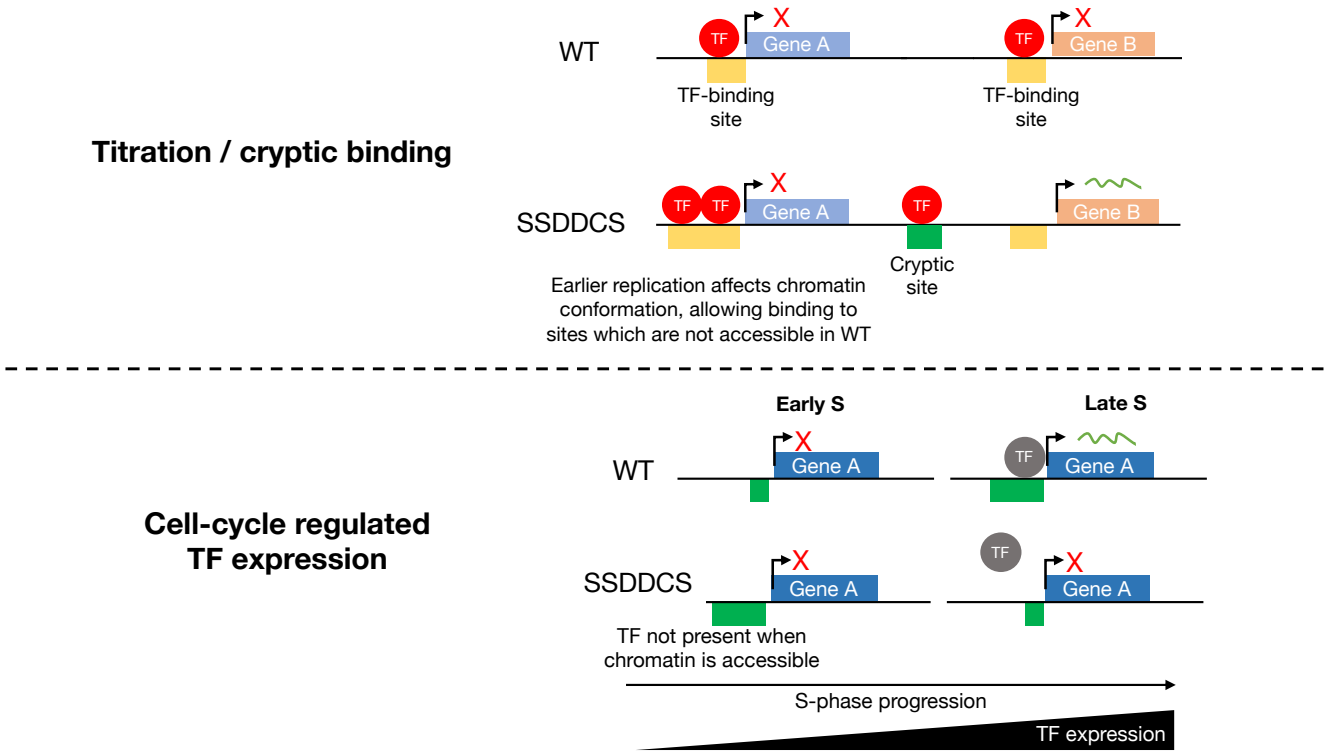


Figure 6.1 – Potential models illustrating the impact of RT on gene expression by perturbing the TF binding landscape. Top – Titration / cryptic binding model. The genome-wide advance in RT caused by SSDDCS overexpression could out-titrate TFs to sites which are more accessible or cryptic sites which are normally not used. In this hypothetical scenario, the TF acts as a repressor of gene expression and silences gene A and gene B in wild-type cells. The overexpression of SSDDCS and consequent RT advance out-titrates this TF from the promoter of gene B to more accessible sites (Gene A) or cryptic sites, allowing expression of gene B**. Bottom – Cell cycle regulation of TF expression.** In this scenario, gene A is replicated and expressed in late S-phase when the promoter is in an accessible state for the binding of the regulator TF, which accumulates during S-phase. Earlier replication caused by SSDDCS overexpression makes the promoter accessible in early S-phase, but the TF has not accumulated to levels that allow efficient binding and induction of expression. By the time that the TF has accumulated, the chromatin is no longer accessible and the gene is not induced.

The lack of a significant number of sub-nucleosomal events supporting these models could be due to the limitations of this approach, some of which were already described. The subMNase-Seq protocol provided an unbiased approach that allowed a genome-wide analysis of sub-nucleosomal binding events upon SSDDCS overexpression. However, it is impossible to determine to which TF each binding event corresponds to using this approach. Moreover, the sequencing reads were distributed throughout all potential events, and a significant amount of the isolated fragments still corresponded to nucleosomal events, leading to a significant portion of sequencing capacity being wasted in these fragments. As such, most likely many lowly expressed TFs or transient binding events were not detected.

I tried to overcome these limitations by setting a threshold of read coverage across all samples to identify high confidence sub-nucleosomal events, in combination with annotation to known TF binding sites in order to increase the probability that the events identified are genuine TF binding events. The increased signal in the Gal80 binding sites validated the analytical approach used (Fig. 5.2B and 5.4). Still, these limitations should be taken into consideration while interpreting the results, which would need to be validated with ChIP-Seq experiments of single TFs. ChIP-Seq would overcome the specificity problem (i.e., which TF is bound?) and the sequencing yield problem, as in the case of ChIP a single TF is immunoprecipitated and as such the majority of sequencing reads will come from regions where this TF is bound. Finally, it should be mentioned that in some cases the absence of a differential peak (both detected using MNase or ChIP) does not necessarily mean that an effect on TF dynamics is not present. For example, it was recently described that Ume6 acts as a platform for the recruitment of both activator and repressor proteins[191] suggesting that Ume6 is always bound to meiotic genes, irrespective of their expression levels.

Another aspect that should be taken into consideration are downstream effects caused by the SSDDCS overexpression, which most likely cause changes in gene expression and chromatin which are independent from the RT advance. For example, this strain accumulates DNA damage markers such as Rad52 foci and γH2A[124] and has decreased growth rates when grown for many generations overexpressing SSDDCS[76]. To avoid potential downstream effects caused by the

overexpression of SSDDCS, replication could be modulated locally, by removing origins (such as the experiments described in Fig. 3.13) or adding extra origins to specific genomic regions in a wild-type background.

6.5 – What's next?

A puzzling observation from this study is the significant number of meiotic genes up-regulated upon a genome-wide advance in RT (Fig. 3.11), which draws an interesting parallel with metazoans where RT changes are associated with gene expression during cellular fate commitments. As such, RT could act as a trigger to induce differentiation events, by affecting the expression of specific genes. It would be interesting to follow this hypothesis in diploid yeast undergoing sporulation: are there RT changes between mitotic and pre-meiotic S-phase which are associated with changes in expression of genes involved in meiosis and sporulation? Blitzblau et al. have found that the same origins are active in mitotic and meiotic S-phase, but with differences in the relative replication timing, as initiation is delayed in most origins in pre-meiotic S-phase[134]. Another potential experiment would be the induction of SSDDCS overexpression in diploids and address if these acquired increased meiotic commitment.

Moreover, if RT patterns are important for cellular fate commitment it is expected that these patterns would be transmitted to daughter cells in subsequent cell cycles. As such, it would be interesting to address if the gene expression and TF binding profiles changes caused by SSDDCS overexpression are maintained in subsequent cell cycles. These experiments would address if specific RT patterns act as epigenetic marks which are transmitted to daughter cells to maintain specific cellular states.

6.6 – Final considerations

Due to the close and complex relationship between RT and transcription, a lot is still left to be found regarding the link between the two. In this study we placed replication in a central position, by conditionally advancing RT in a single cell cycle and have identified loci where RT may be directly impacting gene expression and chromatin. Still, the results presented in this thesis show the complex chain of events caused by a global RT advance, and most likely this advance causes a positive feedback loop where untimely transcription and chromatin disorganisation cause further changes in the genome. In the case of *NDT80,* it is possible that earlier replication affects the TF binding landscape, causing changes in gene expression which are associated with chromatin disorganisation, but we were not able to determine if chromatin disorganisation was a cause or consequence of active transcription. Delaying the replication of this region in an otherwise advanced S-phase was enough to restore gene expression patterns. It would be interesting to address if this delay also restored nucleosome positioning and the Ume6 binding profile, closing the circle between RT and its impact on the genome.

Overall, this work illustrates the importance of a defined order of origin firing and how it might have evolved to maintain gene expression, chromatin and TF binding patterns, placing it as a fundamental aspect of the regulation of the genome structure and function. Moreover, it illustrates the power of budding yeast as an experimental system that allows the manipulation of RT in a cell population both at a global and local level, with the temporal resolution required to dissect the relationship between RT and other genomic events.

References

1. Bell, S. P. & Labib, K. Chromosome Duplication in Saccharomyces cerevisiae. *Genetics* **203**, 1027–67 (2016).
2. Taylor, J. H. Asynchronous Duplication of Chromosomes in Cultured Cells of Chinese Hamster. *J. Biophys. Biochem. Cytol.* **7**, 455–463 (1960).
3. LIMA DE FARIA, A. Incorporation of tritiated thymidine into meiotic chromosomes. *Science* **130**, 503–4 (1959).
4. Raghuraman, M. K. *et al.* Replication Dynamics of the Yeast Genome. *Science (80-.).* **294**, 115–121 (2001).
5. Cayrou, C. *et al.* Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. *Genome Res.* **21**, 1438–1449 (2011).
6. Rhind, N. & Gilbert, D. M. DNA replication timing. *Cold Spring Harb. Perspect. Biol.* **5**, a010132 (2013).
7. Hiratani, I. *et al.* Global Reorganization of Replication Domains During Embryonic Stem Cell Differentiation. *PLoS Biol.* **6**, e245 (2008).
8. Ryba, T. *et al.* Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.* **20**, 761–70 (2010).
9. Donley, N. & Thayer, M. J. DNA replication timing, genome stability and cancer: Late and/or delayed DNA replication timing is associated with increased genomic instability. *Semin. Cancer Biol.* **23**, 80–89 (2013).
10. Stinchcomb, D. T., Struhl, K. & Davis, R. W. Isolation and characterisation of a yeast chromosomal replicator. *Nat. 1979 2825734* **282**, 39–43 (1979).
11. Brewer, B. J. & Fangman, W. L. The localization of replication origins on ARS plasmids in S. cerevisiae. *Cell* **51**, 463–471 (1987).
12. Broach, J. R. *et al.* Localization and Sequence Analysis of Yeast Origins of DNA Replication. *Cold Spring Harb. Symp. Quant. Biol.* **47**, 1165–1173 (1983).
13. Bell, S. P. & Stillman, B. ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. *Nature* **357**, 128–134 (1992).

14. Nieduszynski, C. A., Knox, Y. & Donaldson, A. D. Genome-wide identification of replication origins in yeast by comparative genomics. *Genes Dev.* **20**, 1874 (2006).
15. Zhang, H. & Tower, J. Sequence requirements for function of the Drosophila chorion gene locus ACE3 replicator and ori-beta origin elements. *Development* **131**, 2089–2099 (2004).
16. Segurado, M., de Luis, A. & Antequera, F. Genome-wide distribution of DNA replication origins at A+T-rich islands in Schizosaccharomyces pombe. *EMBO Rep.* **4**, 1048–1053 (2003).
17. Heinzel, S. S., Krysan, P. J., Tran, C. T. & Calos, M. P. Autonomous DNA replication in human cells is affected by the size and the source of the DNA. *Mol. Cell. Biol.* **11**, 2263 (1991).
18. Prioleau, M.-N. & MacAlpine, D. M. DNA replication origins—where do we begin? *Genes Dev.* **30**, 1683–1697 (2016).
19. Meselson, M. & Stahl, F. W. The replication of DNA in Escherichia coli. *Proc. Natl. Acad. Sci.* **44**, 671–682 (1958).
20. McCarroll, R. M. & Fangman, W. L. Time of replication of yeast centromeres and telomeres. *Cell* **54**, 505–513 (1988).
21. Arnoult, N. *et al.* Replication Timing of Human Telomeres Is Chromosome Arm–Specific, Influenced by Subtelomeric Structures and Connected to Nuclear Localization. *PLoS Genet.* **6**, 1000920 (2010).
22. Ten Hagen, K. G., Gilbert, D. M., Willard, H. F. & Cohen, S. N. Replication timing of DNA sequences associated with human centromeres and telomeres. *Mol. Cell. Biol.* **10**, 6348–6355 (1990).
23. Kim, S. M., Dubey, D. D. & Huberman, J. A. Early-replicating heterochromatin. *Genes Dev.* **17**, 330–335 (2003).
24. Yamashita, M. *et al.* The efficiency and timing of initiation of replication of multiple replicons of Saccharomyces cerevisiae chromosome VI. *Genes to Cells* **2**, 655–665 (1997).
25. Friedman, K. L., Brewer, B. J. & Fangman, W. L. Replication profile of Saccharomyces cerevisiae chromosome VI. *Genes to Cells* **2**, 667–678 (1997).
26. Brabant, A. J. van, Hunt, S. Y., Fangman, W. L. & Brewer, B. J. Identifying

sites of replication initiation in yeast chromosomes: Looking for origins in all the right places. *Electrophoresis* **19**, 1239–1246 (1998).

27. Raghuraman, M. K. & Brewer, B. J. Molecular analysis of the replication program in unicellular model organisms. *Chromosom. Res. 2009 181* **18**, 19–34 (2009).
28. Müller, C. A. *et al.* The dynamics of genome replication using deep sequencing. *Nucleic Acids Res.* **42**, e3 (2014).
29. Batrakou, D. G., Müller, C. A., Wilson, R. H. C. & Nieduszynski, C. A. DNA copy-number measurement of genome replication dynamics by high-throughput sequencing: the sort-seq, sync-seq and MFA-seq family. *Nat. Protoc.* **15**, 1255–1284 (2020).
30. Koç, A., Wheeler, L. J., Mathews, C. K. & Merrill, G. F. Hydroxyurea Arrests DNA Replication by a Mechanism That Preserves Basal dNTP Pools. *J. Biol. Chem.* **279**, 223–230 (2004).
31. Feng, W. *et al.* Genomic mapping of single-stranded DNA in hydroxyurea-challenged yeasts identifies origins of replication. *Nat. Cell Biol. 2006 82* **8**, 148–155 (2006).
32. Santocanale, C. & Diffley, J. F. X. A Mec1- and Rad53-dependent checkpoint controls late-firing origins of DNA replication. *Nat. 1998 3956702* **395**, 615–618 (1998).
33. Knott, S. R. V., Viggiani, C. J., Tavare, S. & Aparicio, O. M. Genome-wide replication profiles indicate an expansive role for Rpd3L in regulating replication initiation timing or efficiency, and reveal genomic loci of Rpd3 function in Saccharomyces cerevisiae. *Genes Dev.* **23**, 1077–1090 (2009).
34. Hiratani, I. *et al.* Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome Res.* **20**, 155–69 (2010).
35. Wyrick, J. J. *et al.* Genome-Wide Distribution of ORC and MCM Proteins in S. cerevisiae: High-Resolution Mapping of Replication Origins. *Science (80-.).* **294**, 2357–2360 (2001).
36. Sekedat, M. D. *et al.* GINS motion reveals replication fork progression is remarkably uniform throughout the yeast genome. *Mol. Syst. Biol.* **6**, 353 (2010).

37. Manukyan, A., Abraham, L., Dungrawala, H. & Schneider, B. L. Synchronization of Yeast. in *Methods in molecular biology (Clifton, N.J.)* **761**, 173–200 (2011).
38. Czajkowsky, D. M., Liu, J., Hamlin, J. L. & Shao, Z. DNA Combing Reveals Intrinsic Temporal Disorder in the Replication of Yeast Chromosome VI. *J. Mol. Biol.* **375**, 12–19 (2008).
39. Wang, W. *et al.* Genome-wide mapping of human DNA replication by optical replication mapping supports a stochastic model of eukaryotic replication. *Mol. Cell* **81**, 2975-2988.e6 (2021).
40. Petryk, N. *et al.* Replication landscape of the human genome. *Nat. Commun. 2016 71* **7**, 1–13 (2016).
41. Hawkins, M. *et al.* High-Resolution Replication Profiles Define the Stochastic Nature of Genome Replication Initiation and Termination. *Cell Rep.* **5**, 1132–1141 (2013).
42. Retkute, R., Nieduszynski, C. A. & Moura, A. de. Dynamics of DNA Replication in Yeast. *Phys. Rev. Lett.* **107**, 068103 (2011).
43. Yang, S. C. H., Rhind, N. & Bechhoefer, J. Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing. *Mol. Syst. Biol.* **6**, 404 (2010).
44. Dileep, V. & Gilbert, D. M. Single-cell replication profiling to measure stochastic variation in mammalian replication timing. *Nat. Commun.* **9**, (2018).
45. Donaldson, A. D. & Nieduszynski, C. A. Genome-wide analysis of DNA replication timing in single cells: Yes! We're all individuals. *Genome Biol. 2019 201* **20**, 1–4 (2019).
46. Müller, C. A. *et al.* Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads. *Nat. Methods* (2019). doi:10.1038/s41592-019-0394-y
47. Taylor, J. H. Increase in DNA replication sites in cells held at the beginning of S phase. *Chromosoma* **62**, 291–300 (1977).
48. Eaton, M. L., Galani, K., Kang, S., Bell, S. P. & MacAlpine, D. M. Conserved nucleosome positioning defines replication origins. *Genes Dev.* **24**, 748–53 (2010).

49. Remus, D. *et al.* Concerted Loading of Mcm2–7 Double Hexamers around DNA during DNA Replication Origin Licensing. *Cell* **139**, 719–730 (2009).
50. Miller, T. C. R., Locke, J., Greiwe, J. F., Diffley, J. F. X. & Costa, A. Mechanism of head-to-head MCM double-hexamer formation revealed by cryo-EM. *Nature* **575**, 704 (2019).
51. Randell, J. C. W. *et al.* Mec1 Is One of Multiple Kinases that Prime the Mcm2-7 Helicase for Phosphorylation by Cdc7. *Mol. Cell* **40**, 353–363 (2010).
52. Heller, R. C. *et al.* Eukaryotic Origin-Dependent DNA Replication In Vitro Reveals Sequential Action of DDK and S-CDK Kinases. *Cell* **146**, 80–91 (2011).
53. Zegerman, P. & Diffley, J. F. X. Phosphorylation of Sld2 and Sld3 by cyclin-dependent kinases promotes DNA replication in budding yeast. *Nature* **445**, 281–285 (2007).
54. Muramatsu, S., Hirai, K., Tak, Y.-S., Kamimura, Y. & Araki, H. CDK-dependent complex formation between replication proteins Dpb11, Sld2, Pol , and GINS in budding yeast. *Genes Dev.* **24**, 602–612 (2010).
55. Douglas, M. E., Ali, F. A., Costa, A. & Diffley, J. F. X. The mechanism of eukaryotic CMG helicase activation. *Nature* **555**, (2018).
56. Remus, D. & Diffley, J. F. Eukaryotic DNA replication control: Lock and load, then fire. *Curr. Opin. Cell Biol.* **21**, 771–777 (2009).
57. Nguyen, V. Q., Co, C. & Li, J. J. Cyclin-dependent kinases prevent DNA re-replication through multiple mechanisms. *Nature* **411**, 1068–1074 (2001).
58. Woodward, A. M. *et al.* Excess Mcm2–7 license dormant origins of replication that can be used under conditions of replicative stress. *J. Cell Biol.* **173**, 673 (2006).
59. Rivera-Mulia, J. C. & Gilbert, D. M. Replicating Large Genomes: Divide and Conquer. *Mol. Cell* **62**, 756–765 (2016).
60. Newman, T. J., Mamun, M. A., Nieduszynski, C. A. & Blow, J. J. Replisome stall events have shaped the distribution of replication origins in the genomes of yeasts. *Nucleic Acids Res.* **41**, 9705 (2013).
61. Mamun, M. Al *et al.* Inevitability and containment of replication errors for eukaryotic genome lengths spanning megabase to gigabase. *Proc. Natl.*

Acad. Sci. U. S. A. **113**, E5765–E5774 (2016).

62. Moreno, A. *et al.* Unreplicated DNA remaining from unperturbed S phases passes through mitosis for resolution in daughter cells. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E5757–E5764 (2016).
63. Rhind, N. DNA replication timing: random thoughts about origin firing. *Nat. Cell Biol.* **8**, 1313 (2006).
64. Raghuraman, M. K., Brewer, B. J. & Fangman, W. L. Cell Cycle-Dependent Establishment of a Late Replication Program. *Science (80-.).* **276**, 806–809 (1997).
65. Dimitrova, D. S. & Gilbert, D. M. The spatial position and replication timing of chromosomal domains are both established in early G1 phase. *Mol. Cell* **4**, 983–993 (1999).
66. Belsky, J. A., MacAlpine, H. K., Lubelsky, Y., Hartemink, A. J. & MacAlpine, D. M. Genome-wide chromatin footprinting reveals changes in replication origin architecture induced by pre-RC assembly. *Genes Dev.* **29**, 212–24 (2015).
67. Das, S. P. & Rhind, N. How and Why Multiple MCMs are Loaded at Origins of DNA Replication. *Bioessays* **38**, 613 (2016).
68. DeBeer, M. A. P., Müller, U. & Fox, C. A. Differential DNA affinity specifies roles for the origin recognition complex in budding yeast heterochromatin. *Genes Dev.* **17**, 1817–1822 (2003).
69. Ferguson, B. M. & Fangman, W. L. A position effect on the time of replication origin activation in yeast. *Cell* **68**, 333–9 (1992).
70. Lõoke, M., Kristjuhan, K., Värv, S. & Kristjuhan, A. Chromatin-dependent and -independent regulation of DNA replication origin activation in budding yeast. *EMBO Rep.* **14**, 191–8 (2013).
71. Hoggard, T., Shor, E., Müller, C. A., Nieduszynski, C. A. & Fox, C. A. A Link between ORC-Origin Binding Mechanisms and Origin Activation Time Revealed in Budding Yeast. *PLOS Genet.* **9**, e1003798 (2013).
72. Dukaj, L. & Rhind, N. The capacity of origins to load MCM establishes replication timing patterns. *PLOS Genet.* **17**, e1009467 (2021).
73. Azmi, I. F. *et al.* Nucleosomes influence multiple steps during replication initiation. *Elife* **6**, (2017).

74. Kirstein, N. *et al.* Human ORC/MCM density is low in active genes and correlates with replication time but does not delimit initiation zones. *Elife* **10**, (2021).
75. Foss, E. J. *et al.* Chromosomal Mcm2-7 distribution and the genome replication program in species from yeast to humans. *PLoS Genet.* **17**, (2021).
76. Mantiero, D., Mackenzie, A., Donaldson, A. & Zegerman, P. Limiting replication initiation factors execute the temporal programme of origin firing in budding yeast. *EMBO J.* **30**, 4805–4814 (2011).
77. Tanaka, S., Nakato, R., Katou, Y., Shirahige, K. & Araki, H. Origin association of Sld3, Sld7, and Cdc45 proteins is a key step for determination of origin-firing timing. *Curr. Biol.* **21**, 2055–2063 (2011).
78. Lynch, K. L., Alvino, G. M., Kwan, E. X., Brewer, B. J. & Raghuraman, M. K. The effects of manipulating levels of replication initiation factors on origin firing efficiency in yeast. *PLOS Genet.* **15**, e1008430 (2019).
79. Yoshida, K. *et al.* The Histone Deacetylases Sir2 and Rpd3 Act on Ribosomal DNA to Control the Replication Program in Budding Yeast. *Mol. Cell* **54**, 691–697 (2014).
80. Stevenson, J. B. & Gottschling, D. E. Telomeric chromatin modulates replication timing near chromosome ends. *Genes Dev.* **13**, 146–51 (1999).
81. Natsume, T. *et al.* Kinetochores Coordinate Pericentromeric Cohesion and Early DNA Replication by Cdc7-Dbf4 Kinase Recruitment. *Mol. Cell* **50**, 661–674 (2013).
82. Davé, A., Cooley, C., Garg, M. & Bianchi, A. Protein Phosphatase 1 Recruitment by Rif1 Regulates DNA Replication Origin Firing by Counteracting DDK Activity. *Cell Rep.* **7**, 53–61 (2014).
83. Sreesankar, E., Senthilkumar, R., Bharathi, V., Mishra, R. K. & Mishra, K. Functional diversification of yeast telomere associated protein, Rif1, in higher eukaryotes. *BMC Genomics* **13**, 1–13 (2012).
84. Yamazaki, S. *et al.* Rif1 regulates the replication timing domains on the human genome. *EMBO J.* **31**, 3667–3677 (2012).
85. Buonomo, S. B. C. Rif1-Dependent Regulation of Genome Replication in Mammals. *Adv. Exp. Med. Biol.* **1042**, 259–272 (2017).

86. Knott, S. R. V *et al.* Forkhead transcription factors establish origin timing and long-range clustering in S. cerevisiae. *Cell* **148**, 99–111 (2012).
87. Fang, D. *et al.* Dbf4 recruitment by forkhead transcription factors defines an upstream rate-limiting step in determining origin firing timing. *Genes Dev.* **31**, 2405–2415 (2017).
88. Ostrow, A. Z. *et al.* Conserved forkhead dimerization motif controls DNA replication timing and spatial organization of chromosomes in *S. cerevisiae*. *Proc. Natl. Acad. Sci.* **114**, E2411–E2419 (2017).
89. Gambus, A. *et al.* A key role for Ctf4 in coupling the MCM2-7 helicase to DNA polymerase α within the eukaryotic replisome. *EMBO J.* **28**, 2992–3004 (2009).
90. Jenkinson, F. L. V *et al.* Dephosphorylation of the pre-initiation complex during S-phase is critical for origin firing. *bioRxiv* 2021.11.02.466916 (2021). doi:10.1101/2021.11.02.466916
91. Donaldson, A. D., Fangman, W. L. & Brewer, B. J. Cdc7 is required throughout the yeast S phase to activate replication origins. *Genes Dev.* **12**, 491 (1998).
92. Donaldson, A. D. *et al.* CLB5-Dependent Activation of Late Replication Origins in S. cerevisiae. *Mol. Cell* **2**, 173–182 (1998).
93. Lima-De-Faria, A. & Jaworska, H. Late DNA Synthesis in Heterochromatin. *Nat. 1968 2175124* **217**, 138–142 (1968).
94. Zhang, J., Xu, F., Hashimshony, T., Keshet, I. & Cedar, H. Establishment of transcriptional competence in early and late S phase. *Nat. 2002 4206912* **420**, 198–202 (2002).
95. Woodfine, K. *et al.* Replication timing of the human genome. *Hum. Mol. Genet.* **13**, 191–202 (2004).
96. Gilbert, N. *et al.* Chromatin architecture of the human genome: Gene-rich domains are enriched in open chromatin fibers. *Cell* **118**, 555–566 (2004).
97. Farkash-Amar, S. *et al.* Global organization of replication time zones of the mouse genome. *Genome Res.* **18**, 1562 (2008).
98. Schübeler, D. *et al.* Genome-wide DNA replication profile for Drosophila melanogaster: a link between transcription and replication timing. *Nat. Genet.*

2002 323 **32**, 438–442 (2002).

99. Klein, K. N. *et al.* Replication timing maintains the global epigenetic state in human cells. *Science* **372**, 371–378 (2021).
100. Omberg, L. *et al.* Global effects of DNA replication and DNA replication origin activity on eukaryotic gene expression. *Mol. Syst. Biol.* **5**, 312 (2009).
101. Fraser, H. B. Cell-cycle regulated transcription associates with DNA replication timing in yeast and human. *Genome Biol.* **14**, R111 (2013).
102. Müller, C. A. & Nieduszynski, C. A. DNA replication timing influences gene expression level. *J. Cell Biol.* **216**, 1907–1914 (2017).
103. Hereford, L. M., Osley, M. A., Ludwig, J. R. & McLaughlin, C. S. Cell-cycle regulation of yeast histone mRNA. *Cell* **24**, 367–375 (1981).
104. Couturier, E. & Rocha, E. P. C. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Mol. Microbiol.* **59**, 1506–1518 (2006).
105. Slager, J., Kjos, M., Attaiech, L. & Veening, J. W. Antibiotic-induced replication stress triggers bacterial competence by increasing gene dosage near the origin. *Cell* **157**, 395–406 (2014).
106. Voichek, Y., Bar-Ziv, R. & Barkai, N. Expression homeostasis during DNA replication. *Science (80-.).* **351**, (2016).
107. Padovan-Merhar, O. *et al.* Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol. Cell* **58**, 339–352 (2015).
108. Bell, O. *et al.* Accessibility of the Drosophila genome discriminates PcG repression, H4K16 acetylation and replication timing. *Nat. Struct. Mol. Biol. 2010 177* **17**, 894–900 (2010).
109. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nat. 2014 5157527* **515**, 355–364 (2014).
110. Gindin, Y., Valenzuela, M. S., Aladjem, M. I., Meltzer, P. S. & Bilke, S. A chromatin structure-based model accurately predicts DNA replication timing in human cells. *Mol. Syst. Biol.* **10**, (2014).
111. Pope, B. D. *et al.* Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402–405 (2014).

112. Rivera-Mulia, J. C. & Gilbert, D. M. Replication timing and transcriptional control: beyond cause and effect—part III. *Curr. Opin. Cell Biol.* **40**, 168 (2016).
113. Rivera-Mulia, J. C. & Gilbert, D. M. Replication timing and transcriptional control: Beyond cause and effect - part III. *Curr. Opin. Cell Biol.* **40**, 168–178 (2016).
114. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nat. 2012 4857398* **485**, 376–380 (2012).
115. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
116. Hsieh, T. H. S. *et al.* Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162**, 108–119 (2015).
117. Eser, U. *et al.* Form and function of topologically associating genomic domains in budding yeast. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E3061–E3070 (2017).
118. Lazar-Stefanita, L. *et al.* Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle. *EMBO J.* **36**, 2684–2697 (2017).
119. Ferreira, J. & Carmo-Fonseca, M. Genome replication in early mouse embryos follows a defined temporal and spatial order. *J. Cell Sci.* **110**, 889–897 (1997).
120. Ke, Y. *et al.* 3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis. *Cell* **170**, 367-381.e20 (2017).
121. Oldach, P. & Nieduszynski, C. A. Cohesin-Mediated Genome Architecture Does Not Define DNA Replication Timing Domains. *Genes (Basel).* **10**, (2019).
122. Sima, J. *et al.* Identifying cis Elements for Spatiotemporal Control of Mammalian DNA Replication. *Cell* (2018). doi:10.1016/j.cell.2018.11.036
123. Zegerman, P. & Diffley, J. F. X. Checkpoint-dependent inhibition of DNA replication initiation by Sld3 and Dbf4 phosphorylation. *Nature* **467**, 474–478 (2010).
124. Morafraile, E. C. *et al.* Checkpoint inhibition of origin firing prevents DNA topological stress. *Genes Dev.* (2019). doi:10.1101/gad.328682.119

125. Rivera-Mulia, J. C. *et al.* Cellular senescence induces replication stress with almost no affect on DNA replication timing. *Cell Cycle* **17**, 1667–1681 (2018).
126. Hills, S. A. & Diffley, J. F. X. DNA Replication and Oncogene-Induced Replicative Stress. *Curr. Biol.* **24**, R435–R444 (2014).
127. Koren, A., Soifer, I. & Barkai, N. MRC1-dependent scaling of the budding yeast DNA replication timing program. *Genome Res.* **20**, 781–790 (2010).
128. Takahashi, S. *et al.* Genome-wide stability of the DNA replication program in single mammalian cells. *Nat. Genet.* **51**, 529–540 (2019).
129. Rivera-Mulia, J. C. *et al.* Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells. *Genome Res.* **25**, (2015).
130. Dileep, V. *et al.* Rapid Irreversible Transcriptional Reprogramming in Human Stem Cells Accompanied by Discordance between Replication Timing and Chromatin Compartment. *Stem Cell Reports* (2019). doi:10.1016/J.STEMCR.2019.05.021
131. Rivera-Mulia, J. C. *et al.* Replication timing networks reveal a link between transcription regulatory circuits and replication timing control. *Genome Res.* **29**, 1415–1428 (2019).
132. Mitchell, A. P. Control of meiotic gene expression in Saccharomyces cerevisiae. *Microbiol. Rev.* **58**, 56 (1994).
133. Mori, S. & Shirahige, K. Perturbation of the activity of replication origin by meiosis-specific transcription. *J. Biol. Chem.* **282**, 4447–4452 (2007).
134. Blitzblau, H. G., Chan, C. S., Hochwagen, A. & Bell, S. P. Separation of DNA Replication from the Assembly of Break-Competent Meiotic Chromosomes. *PLoS Genet.* **8**, e1002643 (2012).
135. Takagi, N. Differentiation of X chromosomes in early female mouse embryos. *Exp. Cell Res.* **86**, 127–135 (1974).
136. Epner, E., Forrester, W. C. & Groudine, M. Asynchronous DNA replication within the human beta-globin gene locus. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 8081–8085 (1988).
137. Li, J., Santoro, R., Koberna, K. & Grummt, I. The chromatin remodeling complex NoRC controls replication timing of rRNA genes. *EMBO J.* **24**, 120–

127 (2005).

138. Koren, A. *et al.* Genetic variation in human DNA replication timing. *Cell* **159**, 1015–1026 (2014).
139. Ding, Q. *et al.* The genetic architecture of DNA replication timing in human pluripotent stem cells. *Nat. Commun. 2021 121* **12**, 1–18 (2021).
140. Tadros, W. & Lipshitz, H. D. The maternal-to-zygotic transition: a play in two acts. *Development* **136**, 3033–3042 (2009).
141. Seller, C. A. & O'Farrell, P. H. Rif1 prolongs the embryonic S phase at the Drosophila mid-blastula transition. *PLoS Biol.* **16**, (2018).
142. Siefert, J. C., Georgescu, C., Wren, J. D., Koren, A. & Sansam, C. L. DNA replication timing during development anticipates transcriptional programs and parallels enhancer activation. *Genome Res.* **27**, 1406–1416 (2017).
143. Pourkarimi, E., Bellush, J. M. & Whitehouse, I. Spatiotemporal coupling and decoupling of gene transcription with DNA replication origins during embryogenesis in C. elegans. *Elife* **5**, (2016).
144. Collart, C., Allen, G. E., Bradshaw, C. R., Smith, J. C. & Zegerman, P. Titration of Four Replication Factors Is Essential for the Xenopus laevis Midblastula Transition. *Science (80-.).* **341**, 893–896 (2013).
145. Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393 (2009).
146. Herrick, J. Genetic variation and DNA replication timing, or why is there late replicating DNA? *Evolution* **65**, 3031–3047 (2011).
147. Blumenfeld, B., Ben-Zimra, M. & Simon, I. Perturbations in the Replication Program Contribute to Genomic Instability in Cancer. *Int. J. Mol. Sci.* **18**, (2017).
148. Lang, G. I. & Murray, A. W. Mutation Rates across Budding Yeast Chromosome VI Are Correlated with Replication Timing. *Genome Biol. Evol.* **3**, 799–811 (2011).
149. Taxis, C. *et al.* Spore number control and breeding in Saccharomyces cerevisiae: a key role for a self-organizing system. *J. Cell Biol.* **171**, 627 (2005).
150. De, S. & Michor, F. DNA replication timing and long-range DNA interactions

predict mutational landscapes of cancer genomes. *Nat. Biotechnol. 2011 2912* **29**, 1103–1108 (2011).

151. Hiratani, I. & Gilbert, D. M. Replication timing as an epigenetic mark. *Epigenetics* **4**, 93 (2009).
152. Ryba, T. *et al.* Abnormal developmental control of replication-timing domains in pediatric acute lymphoblastic leukemia. *Genome Res.* **22**, 1833–1844 (2012).
153. Du, Q. *et al.* Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. *Nat. Commun. 2019 101* **10**, 1–15 (2019).
154. Courtot, L. *et al.* Low replication stress leads to specific replication timing advances associated to chromatin remodelling in cancer cells. *bioRxiv* 2020.08.19.256883 (2020). doi:10.1101/2020.08.19.256883
155. Nocetti, N. & Whitehouse, I. Nucleosome repositioning underlies dynamic gene expression. *Genes Dev.* **30**, 660–672 (2016).
156. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
157. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
158. Balakrishnan, R. *et al.* YeastMine--an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. *Database (Oxford).* **2012**, bar062 (2012).
159. Chen, K. *et al.* DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res.* **23**, 341–51 (2013).
160. Gutiérrez, M. P., MacAlpine, H. K. & MacAlpine, D. M. Nascent chromatin occupancy profiling reveals locus- and factor-specific chromatin maturation dynamics behind the DNA replication fork. *Genome Res.* **29**, 1123–1133 (2019).
161. De Boer, C. G. & Hughes, T. R. YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Res.* **40**, D169–D179 (2012).

162. Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91–D94 (2004).
163. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
164. Briu, L. M., Maric, C. & Cadoret, J. C. Replication Stress, Genomic Instability, and Replication Timing: A Complex Relationship. *Int. J. Mol. Sci. 2021, Vol. 22, Page 4764* **22**, 4764 (2021).
165. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS One* **6**, e21800 (2011).
166. Rachidi, N., Martinez, M. J., Barre, P. & Blondin, B. Saccharomyces cerevisiae PAU genes are induced by anaerobiosis. *Mol. Microbiol.* **35**, 1421–1430 (2000).
167. Wightman, R. & Meacock, P. A. The THI5 gene family of Saccharomyces cerevisiae: distribution of homologues among the hemiascomycetes and functional redundancy in the aerobic biosynthesis of thiamin from pyridoxine. *Microbiology* **149**, 1447–1460 (2003).
168. Li, M. *et al.* Thiamine Biosynthesis in Saccharomyces cerevisiae Is Regulated by the NAD+-Dependent Histone Deacetylase Hst1. *Mol. Cell. Biol.* **30**, 3329 (2010).
169. Herskowitz, I. Life cycle of the budding yeast Saccharomyces cerevisiae. *Microbiol. Rev.* **52**, 536 (1988).
170. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705 (1998).
171. Gurevich, V. & Kassir, Y. A Switch from a Gradient to a Threshold Mode in the Regulation of a Transcriptional Cascade Promotes Robust Execution of Meiosis in Budding Yeast. *PLoS One* **5**, e11005 (2010).
172. Jiang, C. & Pugh, B. F. A compiled and systematic reference map of nucleosome positions across the Saccharomyces cerevisiae genome. *Genome Biol.* **10**, R109 (2009).
173. Park, H. D., Luche, R. M. & Cooper, T. G. The yeast UME6 gene product is required for transcriptional repression mediated by the CAR1 URS1 repressor

binding site. *Nucleic Acids Res.* **20**, 1909 (1992).

174. Pnueli, L., Edry, I., Cohen, M. & Kassir, Y. Glucose and Nitrogen Regulate the Switch from Histone Deacetylation to Acetylation for Expression of Early Meiosis-Specific Genes in Budding Yeast. *Mol. Cell. Biol.* **24**, 5197 (2004).
175. Strich, R., Slater, M. R. & Esposito, R. E. Identification of negative regulatory genes that govern the expression of early meiotic genes in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 10018–10022 (1989).
176. Mallory, M. J., Cooper, K. F. & Strich, R. Meiosis-specific destruction of the Ume6p repressor by the Cdc20-directed APC/C. *Mol. Cell* **27**, 951–961 (2007).
177. Fleming, A. B., Beggs, S., Church, M., Tsukihashi, Y. & Pennings, S. The yeast Cyc8-Tup1 complex cooperates with Hda1p and Rpd3p histone deacetylases to robustly repress transcription of the subtelomeric FLO1 gene. *Biochim. Biophys. Acta* **1839**, 1242–1255 (2014).
178. Tam, J. & van Werven, F. J. Regulated repression governs the cell fate promoter controlling yeast meiosis. *Nat. Commun.* **11**, (2020).
179. Lohr, D., Venkov, P. & Zlatanova, J. Transcriptional regulation in the yeast GAL gene family: a complex genetic network. *FASEB J.* **9**, 777–787 (1995).
180. MacIsaac, K. D. *et al.* An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* **7**, (2006).
181. Zhang, Z. & Reese, J. C. Molecular Genetic Analysis of the Yeast Repressor Rfx1/Crt1 Reveals a Novel Two-Step Regulatory Mechanism. *Mol. Cell. Biol.* **25**, 7399–7411 (2005).
182. Xie, J. *et al.* Sum1 and Hst1 repress middle sporulation-specific gene expression during mitosis in Saccharomyces cerevisiae. *EMBO J.* **18**, 6448–6454 (1999).
183. Kasten, M. M. & Stillman, D. J. Identification of the Saccharomyces cerevisiae genes STB1-STB5 encoding Sin3p binding proteins. *Mol. Gen. Genet.* **256**, 376–386 (1997).
184. Sbia, M. *et al.* Regulation of the yeast Ace2 transcription factor during the cell cycle. *J. Biol. Chem.* **283**, 11135–11145 (2008).
185. Ishihara, M. *et al.* Protein phosphatase type 1-interacting protein Ysw1 is

involved in proper septin organization and prospore membrane formation during sporulation. *Eukaryot. Cell* **8**, 1027–1037 (2009).

186. Lin, C. P. C., Kim, C., Smith, S. O. & Neiman, A. M. A highly redundant gene network controls assembly of the outer spore wall in S. cerevisiae. *PLoS Genet.* **9**, (2013).
187. Lenstra, T. L. *et al.* The Specificity and Topology of Chromatin Interaction Pathways in Yeast. *Mol. Cell* **42**, 536–549 (2011).
188. Hu, Z. *et al.* Nucleosome loss leads to global transcriptional up-regulation and genomic instability during yeast aging. *Genes Dev.* **28**, 396–408 (2014).
189. Huang, S., Zhou, H., Tarara, J. & Zhang, Z. A novel role for histone chaperones CAF-1 and Rtt106p in heterochromatin silencing. *EMBO J.* **26**, 2274 (2007).
190. McInerny, C. J. Cell Cycle Regulated Gene Expression in Yeasts. *Adv. Genet.* **73**, 51–85 (2011).
191. Raithatha, S. A., Vaza, S., Islam, M. T., Greenwood, B. & Stuart, D. T. Ume6 Acts as a Stable Platform To Coordinate Repression and Activation of Early Meiosis-Specific Genes in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **41**, (2021).