Université de Paris

École doctorale bioSPC – ED562

Institut Jacques Monod – *UMR 7592*

Molecular investigation of *cis*-regulatory elements and chromatin organisation of vertebrate replication origins

Étude moléculaire des éléments *cis*-régulateurs et de l'organisation de la chromatine des origines de réplication chez les vertébrés

---

Par Jérémy POULET- -BENEDETTI

Thèse de doctorat de Biologie moléculaire

Dirigée par Marie-Noëlle PRIOLEAU

Présentée et soutenue publiquement le 15 Septembre 2020

Devant un jury composé de :

Pr. Reiner A. VEITIA (Institut Jacques Monod)	Président du jury
Dr. Maria GOMEZ (Centro de Biologia Molecular « Severo Ochoa »)	Rapporteur
Dr. Julian SALE (MRC Laboratory of Molecular biology)	Rapporteur
Dr. Patricia KANNOUCHE (Institut Gustave Roussy)	Examinateur
Dr. Philippe PASERO (Institut de génétique Humaine)	Examinateur
Dr. Marie-Noëlle PRIOLEAU (Institut Jacques Monod)	Directrice de thèse

« À vaincre sans péril, on triomphe sans gloire. »

Pierre Corneille, *Le cid*.

Acknowledgements

I want to thank the members of my PhD defense committee, Dr. Maria Gomez, Dr. Patricia Kannouche, Dr. Philippe Pasero, Dr. Julian Sale and Pr. Reiner Veitia for their time and expertise for the evaluation of my PhD work. I would like to thank in particular Dr. Maria Gomez and Dr. Julian Sale for being rapporteur of my thesis manuscript. Thank you Dr. Patricia Kannouche and Dr. Philippe Pasero for being examiners of my PhD and thank you Pr. Reiner Veitia for being the jury president.

Je souhaite, en premier lieu, remercier ma directrice de thèse Marie-Noëlle Prioleau, qui m'a permis de réaliser ma thèse dans son laboratoire et m'a poussé dans mes réflexions et dans mon travail. Je souhaite aussi remercier Marie-Noëlle de m'avoir laissé la liberté de me confronter aux problèmes que j'ai rencontrés au cours de mon projet, afin de me laisser évoluer et devenir le scientifique que je suis maintenant. Je la remercie aussi pour sa confiance et pour m'avoir permis de voyager et rencontrer la communauté scientifique de la réplication aux cours de congrès. Je la remercie enfin pour toutes ses corrections lors de l'écriture de ma thèse et je souhaite m'excuser pour mon orthographe plus qu'approximative.

Je souhaite particulièrement remercier Caroline Brossas, qui m'a soutenu tout au long de ma thèse, dans les bons moments et les moments plus difficiles. Merci, pour le temps que tu as pris pour m'apprendre tout ce que tu connaissais et pour ta patience lors de mes « blocages » sur des questions parfois simples. Au même titre que Marie-Noëlle, tu as participé à former le scientifique que je suis maintenant. Je te remercie pour tous les souvenirs que je garderai et pour notre pratique intensive de course à pieds (même si j'étais loin d'être constant !).

Je tiens à remercier Bénédicte Duriez pour toutes les discussions scientifiques, mais aussi pour les discussions culturelles de fin de journée et pour les échanges de livres, de conseils de films et musiques.

Je souhaite remercier Natalja Barinova, pour son aide et ses commentaires tout au long de ma thèse. Je la remercie, ainsi que Caroline, pour nos journées intenses d'élutriation qui étaient du sport à part entière.

J'aimerais remercier toutes les personnes qui sont passées au laboratoire et avec qui j'ai eu l'occasion d'échanger : Nikos Parisis, avec qui j'ai pu discuter du monde de la science, Eslande Hercul, pour nos conversations par bouteilles interposées et nos fous rires, Coralie Goncalves, qui a réussi à râler même en étant de bonne humeur et Alexandra Da Silva Babinet, qui fut un bol d'air frais tout au long de son stag… apprentissage (pardon) et qui a fini par aimer les plantes. Je souhaite remercier Aurore Champigny pour son analyse bio-informatique qui aurait, sans doute, été douloureuse sans elle. Merci,

à tous ceux qui sont passés au laboratoire sans être nommés mais que j'ai apprécié (ou encadré). Je tiens à remercier, Magali Fradet, de la plateforme de cytométrie en flux de l'Institut, qui m'a beaucoup appris en cytométrie en flux et avec qui j'ai beaucoup échangé.

Je souhaite faire une mention spéciale à tout le 5ème étage de l'institut Jacques Monod et plus particulièrement Anne-Laure Todeschini, Vanessa Ribes, May Penrad, Bérangère Legois mais aussi tous les autres pour le partage, les discussions et la bonne ambiance générale.

Je tiens à remercier le « gang » du 5ème étage, Laeti, Line, Noémie, Nico et Youcef. Je ne pourrai pas résumer toutes nos discussions, nos échanges, nos partages, nos souvenirs et tous les bons moments. Chacun a pu compter sur les uns et les autres dans les moments difficiles et sans vous ma thèse aurait été surement un long moment difficile à passer ! Laeti, ma complice des débuts de thèse, Line, qui nous a toujours mis en retard au ciné (sans le faire exprès ?), Noémie à qui il fallait tout répéter deux fois, Nico, avec qui j'ai pu parler « science et PCR » dans les couloirs et Youcef, pour ses courses effrénées au 5ème étage. Alors je dirais seulement, merci pour tout…

Je remercie Norry et Jonathan, qui ont rendu les courses funs et merci pour les à-côtés et l'ambiance générale.

Je souhaite faire un petit clin d'œil à tous ceux que je ne cite pas forcément maintenant, mais que j'ai pu rencontrer au cours de ma thèse, à tous les nouveaux arrivants que j'aurais aimé avoir l'occasion de plus côtoyer, mais qui seront, j'en suis sûr, de bons scientifiques. À tous ceux aussi qui m'ont aperçu à la bibliothèque pendant l'écriture de ma thèse et avec qui j'ai pu discuter.

Bien évidemment, je souhaite remercier ma famille, qui m'a toujours soutenu, durant mes études et dans ma vie, sans qui je n'aurais pas pu poursuivre mes études jusqu'au doctorat et réaliser des stages à l'étranger. Je les remercie pour être ceux qu'ils sont et pour m'avoir aidé à devenir qui je suis maintenant. J'aurais aimé avoir ma grand-mère encore présente pour qu'elle puisse voir son petit-fils devenir docteur.

Enfin, je remercie ma poupi, qui m'a supporté pendant mon doctorat et avec qui j'espère pouvoir faire un bon bout de chemin dans la vie.

Table of contents

Table of Abbreviations

ABF1: ARS-binding factor 1

ACS: Autonomous replicating consensus sequence

ARS: Autonomous replicating sequence

CDC6: Cell division cycle 6

CDK: Cyclin dependent kinases

CDT1: CDC-10 dependent transcript

CGI: CpG Islands

ChIP: Chromatin immunoprecipitation

DDK: DBF4-dependent kinase

DHS: DNAseI hypersensitive site

dNTP: desoxy-nucleoside triphosphate

EtoL: Early to Late Replicating

EWS: Ewing's sarcoma protein

FISH: Fluorescent In Situ Hybridasation

Fkh: Forkhead

G4: G-quadruplex

Ini-seq: initiation sequencing

LAD: lamina associated domains

LtoE: Late to Early Replicating

LYSC: Lysozyme C gene

MCM: Mini-chromosome maintenance

mES: mouse Embryonic Stem cells

mESCs: mouse Embryonic stem cells

MNase: Micrococcal nuclease

NDR: Nucleosome Depleted Region

NF-1: Nuclear factor 1

NFR: nucleosome Free Region

nPCs: neural precursor cells

nr-ACS: non replicative ACS

OCCM: ORC-CDC6-CDT1-MCM

OCM: ORC-CDC6-MCM

OGRE: Origin G-rich repeated elements

Ok-seq: Okazaki sequencing

ORC: Origin recognition complex

PDS: Pyridostatin

pG4: potential G-quadruplex

PP1: Protein phosphatase 1

Pre-IC: Pre initiation complex

Pre-RC: Pre-replication complex

RAD: RIF1 associated domains

RPA: replication Protein A

RT: Replication Timing

SNS: short Nascent Strand

TAD: topological associated domains

TBP: TATA-binding protein

Tm: temperature melting

TSS: Transcription start site

TTR: timing transition regions

Part1
Introduction

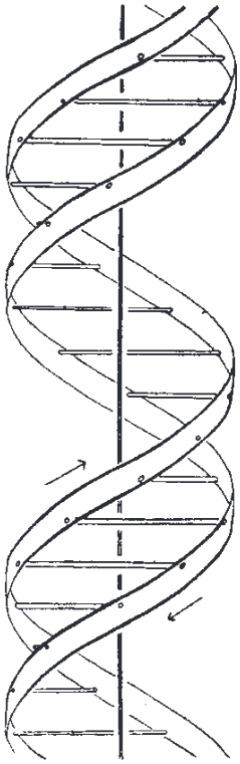


Figure 1:

Schematic representation of anti-parallel DNA double helix, spinning around the axis

From Watson and Crick (1953)

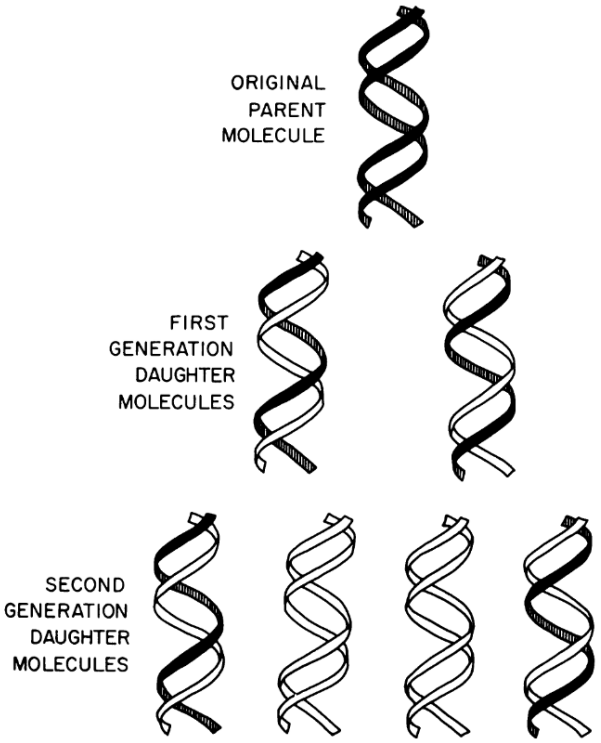


Figure 2

Illustration of the semi conservative DNA transmission from parental molecules to daughter molecules. Black strands represent the mother molecule.

From Meselson and Stahl (1958)

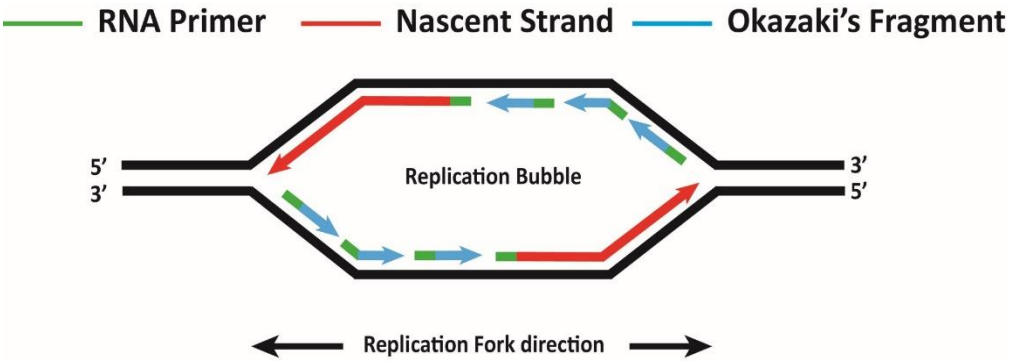


Figure 3

Simplified representation of a replication bubble, with the organization of the newly synthesized DNA at the replication bubble. Arrows represent the DNA molecule orientations starting from 5' to 3'.

1. DNA and Replication

a. A short history

The structure of DNA was firstly described in 1953[1] (figure 1) nearly ten years after the identification and purification of a transforming agent that could produce heritable change in an organism[2] that was identified eleven years later as desoxyribonucleic acid molecules[3]. Those discoveries raised the fundamental question of how DNA could be transferred from one cell to two daughter cells. At this time, three models were competing to answer this problematic, the conservative, the semi-conservative and the dispersive DNA transmission model. Dr Meselson and Dr Stahl made the experiments that solved this issue in 1958. They used Escherichia Coli bacteria and DNA incorporation of isotopes to differentiate between the different models proposed and concluded on the use of the semi-conservative model during cell division[4] (figure 2).

DNA polymerases encounter several challenges during DNA replication. Due to their need of a RNA primer to start DNA synthesis and their 5' to 3' synthesis on an anti-parallel template DNA molecule, DNA replication could not be simply the production of a unique newly synthesised strand. Indeed, after the opening of the DNA molecule and fork progression, opened DNA, from the 3' to the 5' end, can be continuously replicated by DNA polymerases after the synthesis of a small RNA primer. This one long DNA molecule is called leading strand or also Nascent Strand. On the other strand, the opening of the DNA from the 5' to the 3' imposes the synthesis, by DNA polymerases, of several small DNA fragments and to ligate them during the whole replication process. Those small fragments, of around 200bp, are called Okazaki fragments and are synthetized on the lagging strand. DNA polymerases need the presence of a RNA primer to start DNA replication, these primers are deposited by an enzyme called primase. This RNA primer is transiently incorporated to neo-synthesised DNA to later on be replaced by DNA. This RNA primer at the 5' part of DNA is characteristic of intermediate molecules of DNA replication (figure 3).

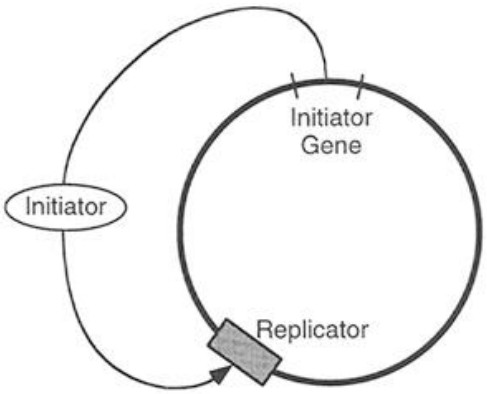


Figure 4

Scheme of the replicon model proposing the expression of the initiator protein recognizing and activating the replicator sequence.

From Jacob and Brenner (1963)

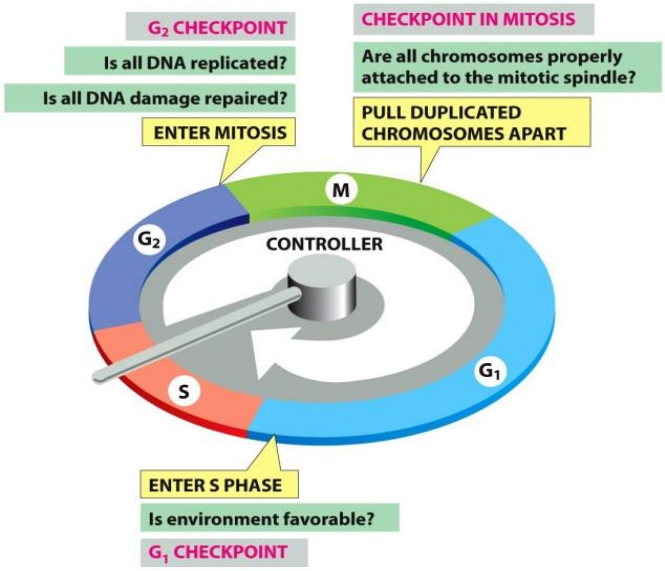
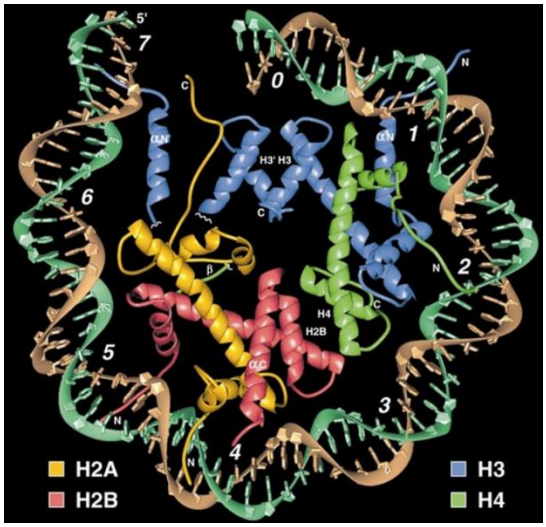


Figure 5

Cartoon of the cell cycle, with the different phases of the cycle. Green boxes represents the assessment of the cells to ensure the good mitosis of the cells.

From molecular biology of the cell, Alberts, Raff, Lewis, Roberts, Bray and Watson

A



B

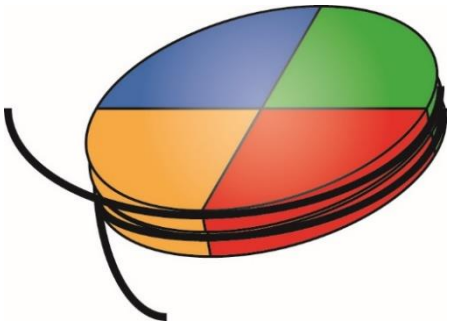


Figure 6

A. Crystal structure of half nucleosome cut vertically, green and light brown strands represent single strand specific DNA molecules. Histone subunits are represented by specific colors, only a partial view of the histone tails can be observed. From luger *et al* (1997) B. Cartoon of the nucleosome organization with two turns of DNA around the histone core.

b. The replicon model

Although, these pioneering experiments identified the molecules carrying the genetic information, its structure and its transmission, they did not address the question of how does replication start. The first model has been proposed in 1963, using again bacteria as a model system, it appeared that only certain sequences were able to support plasmid replication and transmission through cell divisions[5]. This model relies on the recognition of a replicator sequence (*cis*-element) by a *trans*-activating factor called initiator. The initiator is required to destabilise the stable DNA double helix and to allow helicases and DNA polymerases to replicate DNA (figure 4). Such model has been the basis of DNA replication study field and appears to be still up to date.

c. Cell cycle and DNA replication

The molecular mechanisms and structure described earlier have to be put into the context of the cell cycle. The cell cycle is divided in four steps G1, S, G2 and M, which are essential to ensure a proper cell division. The S-Phase (standing for DNA-synthesis) corresponds to DNA replication and the M-phase (for Mitosis) corresponds to the chromosomes segregation. G1 and G2 are cell preparation phases ensuring the cell ability to manage the S and M-phases (figure 5). Checkpoints control the cell cycle phases progression through the activity of Cyclin-dependent kinases (CDK). Tightly regulated cyclins activate the serine/threonine protein kinases CDKs. CDKs ensure phosphorylation of key target proteins that allow the cell cycle progression. Checkpoints can block cell cycle in case of too high DNA damage level , un-replicated DNA or mitotic spindle un-anchoring[6].

d. Chromatin structure inside the nucleus

In eukaryotes, association of DNA with histones proteins into particles named nucleosomes allows genomic DNA to fit inside the nucleus. Histones are relatively small basic proteins (11-15kDa)[7], four sub-units H2A-H2B-H3 and H4 can form an octamer consisting of two copies each[8]. The nucleosome core particle contains about 147 bp of DNA wrapped around histone octamer (Figure 6). Adjacent nucleosomes are joined by a stretch of free DNA termed linker DNA, which varies from 10-80 bp in length depending on species and tissue types. Histones proteins, even though they are associated with DNA, present histone “tails”, that is an un-structured domain of histones. These tails represent 20-30% of the total mass histones cores[8].

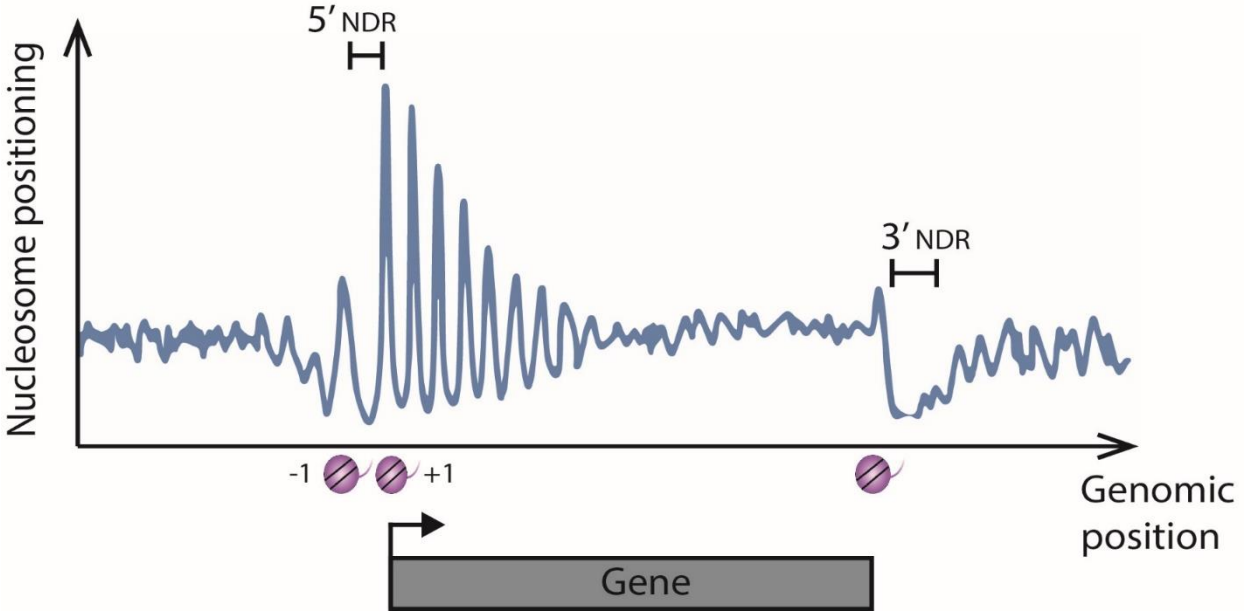


Figure 7

General representation of nucleosome positioning at the gene level showing Nucleosome Depleted Regions (NDR). Purple histones are placed at protected DNA against nuclease digestion. Numbers represent the relative positioning of the nucleosome relative to the TSS.

Adapted from Arya *et al*, (2010)

Those tails, through the activity of histone modifier enzymes, can be post-translationally modified (acetylated, methylated, sumoylated …). Such modifications affect histone-DNA and inter-histones-histones interactions, impacting on nucleosomes stability, secondary and tertiary chromatin organisation[7]. Nucleosomes are a physical barrier for DNA interacting factors. Indeed, DNA associated with histone proteins is not easily accessible. However, nucleosomes can be removed or displaced from DNA by specific factors. Nucleosome Free Regions (NFR)/Nucleosome depleted regions (NDR) can be observed at functionally active sequences, NFR allow *trans*-activating factors to easily reach *cis*-regulating elements. For example, at transcription start sites (TSS) of active genes, an NFR can be observed upstream of the TSS (figure 7). Moreover, nucleosomes can be precisely and strongly positioned, due to sequence characteristics and protein factor constraints[9].

Certain enzymes and factors can recognise specific histone tail modifications. These properties of histone tails gave rise to the histone code hypothesis[10]. Another level of regulation relies on the presence of histones sub-units variants, such variants being involved in different molecular processes. Canonical histones proteins are expressed and deposited on DNA during the S-phase, variants histone proteins are, on contrary, expressed all along the cell cycles and are incorporated to nucleosomes in a DNA replication-independent way[11]. H2A.X has been found enriched at the DNA double strand breaks sites[11]. It has also been observed that H2A.Z and H3.3 histone variants are found enriched at NFR, those variants destabilised the nucleosome resulting in the presence of a labile nucleosome[12].

The histone H1 sub-unit allows the compaction to a more or less dense state of chromatin through interactions with linker DNA and nucleosomal core particles. Open chromatin is called euchromatin and is associated with active transcription, on the other hand, closed chromatin is called heterochromatin and is associated with silenced genes.

e. Molecular basis of replication origin recognition

DNA replication starts at specific sites named replication origins (Replicator according to the replicon model). Origins are activated in a two steps process during the cell cycle. In G1, the Pre-Replication Complex (Pre-RC) is loaded onto origins, this complex is constituted of the Origin Recognition Complex (ORC), Cell Division Cycle 6 (CDC6), CDC10-Dependent Transcript 1 (CDT1) and a Mini-Chromosome Maintenance double hexamer (MCM). Pre-RC loading on replication origins, is called Licensing. Once cells enter the S-phase, only a subset of Pre-RCs are activated through their phosphorylation by CDKs and DBF4-Dependent Kinase (DDK) and by the recruitment of several protein factors such as MCM10, RECQL4, Treslin, CDC45, GINS, TOPB1, DNA polymerase (here in mammals) … Activated origins are also defined as fired origin. The different factors of the Pre-RC are conserved among eukaryotes[13].

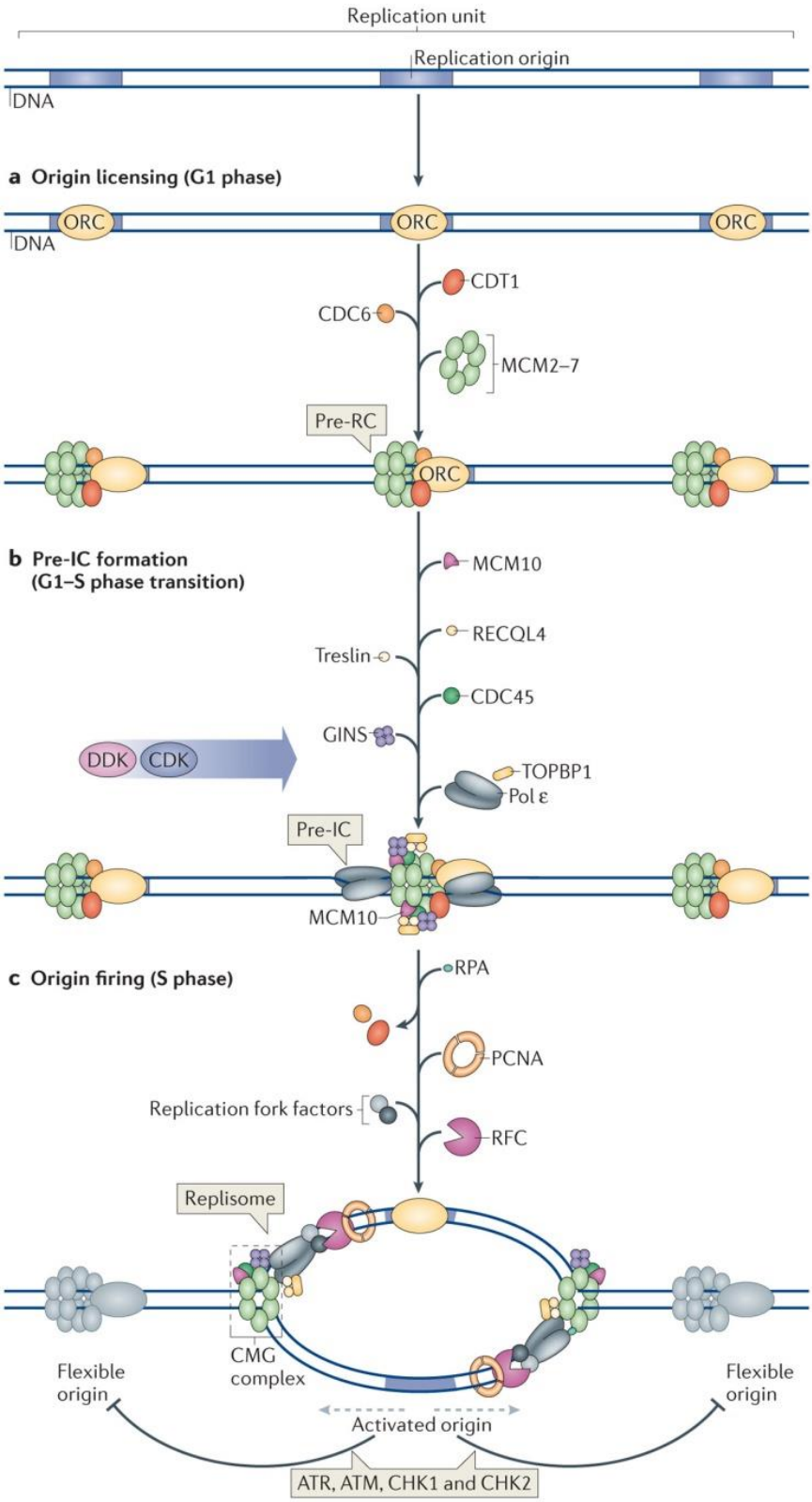


Figure 8

Cartoon of the Pre-RC formation during the G1 phase, followed by the Pre-IC formation and the origin firing in S phase.

From Fragkos *et al*, (2015)

In the yeast *Saccharomyces Cerevisiae*, origins exhibit a consensus sequence called ARS consensus sequence (ACS) this 11 bp AT-rich motif is firstly recognised by the ORC complex[14]. Expressed all along the cell cycle, ORC is constituted of six subunits (ORC1-6) and exhibit AAA+ domains on the ORC1 and ORC5 subunits. In addition, ORC4 protein contain an essential arginine finger that is required for ORC ATPase activity[15]. ORC bound DNA in a cell-cycle independent fashion[16] in a low specificity manner and required CDC6 interaction to ensure its proper anchoring to ACS. CDC6 is another AAA+ protein, tightly regulated and expressed during late M, it interacts with ORC during the G1 phase, inducing a strong ATP hydrolysis activity acting on the OCR/CDC6 complex stability. This ATP hydrolysis activity is ensured at the same time by ORC and CDC6, but only CDC6 ATPase activity acts on the complex stability[17]. Non-origin DNA will promote CDC6 ATP hydrolysis activity inducing ORC/CDC6 complex dissociation, on the other hand, ACS DNA recognized by the ORC/CDC6 complex decrease the ATPase activity, stabilising the complex on the origin[17] (Figure 8).

Pre-RC remaining components, CDT1 and MCM2-7, are localised in the cytoplasm and associated with each other. CDT1 and MCM2-7 are gradually excluded, at the beginning of the S-phase, from the nucleus, to only be imported back at the end of the M-phase[16], this specific cell localisation has been observed only in S.*Cerevisiae*. In mammals Geminin inhibits CDT1 all along the cell cycle, except during G1, when CDT1 is require for Pre-RC licensing[18]. MCM is an AAA+ hexamer complex that act as the replicative helicase at the fork. MCM2-7 association with CDT1 in the cytoplasm induces its relocalisation inside the nucleus. CDT1 interacts with MCM6 C-terminus domain that, when free, blocks MCM2-7 interaction with ORC/CDC6. CDT1 alleviates MCM6 C-Ter inhibitory activity and stabilises the ORC-CDC6-CDT1-MCM2-7(OCCM) complex formation[19]. MCM2-7 is recruited at the ORC/CDC6 complex by the MCM3 C-Terminal domain[20]. Once the OCCM complex is formed, CDC6 and ORC1 ATP hydrolysis induce the release of CDT1 from the OCCM, forming an ORC-CDC6-MCM2-7 (OCM) complex. The OCM allows the loading of the ring-shaped MCM2-7 on the DNA in a ORC6 dependent manner[19]. The second MCM loading requires the OCM CDC6 protein release, leaving the ORC-MCM complex on the DNA. A second ORC complex is recruited to the MCM N-terminal domain through the ORC6 N-terminal domain interaction, allowing at the same time, or before second ORC binding, the first ORC proteins release[21]. The newly bound ORC will recruit CDC6 and, in a CDC6-dependent manner, will allow the loading of a second MCM complex in a head-to-head conformation acting via DNA bending[21] (Figure 8).

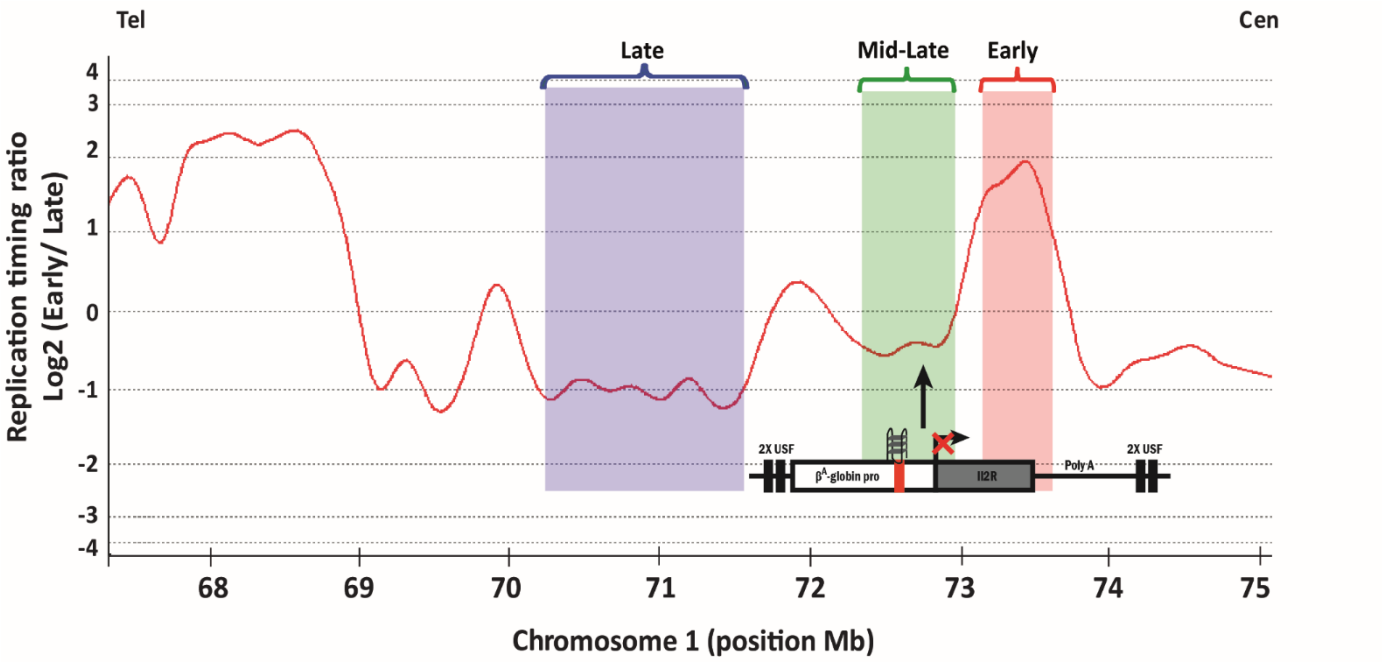


Figure 9

Replication timing program on a part of the chicken chromosome 1 from 67 to 75 Mb. The blue, green and red rectangles represent late, mid-late and early replicating regions respectively. Black arrow: Insertion site. Colored Arrow represent the potential timing shift resulting from ectopic origin insertion.

Pre-RC activation requires CDKs and DDK activity that phosphorylate MCM sub-units, phosphorylated MCM triggers the recruitment of CDC45 and GINS, forming, at the end, the Pre-IC (pre-initiation complex). This is only after the recruitment of several factors such as Replication Protein A (RPA), topoisomerases, DNA polymerase Epsilon and DNA polymerase Alpha that the replication origin can fire and start DNA replication[13,22]. This molecular process has been reconstituted and could support DNA replication *in vitro* with *S. Cerevisiae* 16 purified proteins[23]. These experiments confirmed, *in vitro,* the role of CDKs and DDK for factors recruitments, but also their role in dissociating origin licensing and firing. Indeed, ORC phosphorylation, inhibits its ability to recruit MCM to replication origin, coupled with the degradation of CDC6 and cytoplasm export of CDT1-MCM. This system avoids re-replication that could lead to genomic instability.

2. Replication timing program

DNA replication is a complex mechanism that have to be well organised in order to ensure the good DNA transmission from one cell to the two daughters cells. In bacteria, DNA replication is relatively easy and in most cases relies on the activation of a single replication origin. However, in eukaryotes, the presence of several chromosomes and the increase size of their genomes lead to a high number of origins and the establishment of a very strict replication program organisation. Cells do not have the required amount of factors to ensure the firing of all the replication origins at the same time.

Careful study of newly synthesized DNA during the S-phase revealed that genomic DNA is replicated according to a precise replication timing program that can be divided into different phases[13] (Figure 9). Basically, regions can be divided into three types, early, mid and late-replicating regions corresponding to regions replicated at the beginning, the middle and the end of S-phase respectively. Replication-timing program depiction genome-wide reveals its organisation into large domains of similar timing ranging from several hundreds of kilo-bases to mega-bases.

Molecular mechanisms establishing replication-timing programs are still mostly unknown. Several studies try to characterise and identify potential regulators such as the cis-regulatory elements, trans-factors or chromatin-environment. Most of the studies found correlation but direct causal effects remain to be demonstrated.

Genomic analyses of the timing program associate early-replicating domains with GC-richness as well as gene richness and the late-replicating region with GC-poor and gene poor sequences[13]. Genes located inside early replicating regions are mostly housekeeping genes and genes located in late regions are mostly tissue-specific. Moreover, the establishment of genome-wide origin maps show

that early replication timing domains are dense in efficient origins compared to late replicating regions that are origin poor (see section 4 Replication origins in eukaryotes).

Timing transition regions (TTR) that are the regions connecting early to late replication-timing domains have first been proposed to be regions devoid of replication origins, replicated by a unique unidirectional fork, emerging from the early domain to the later domain. However, such model relies on the use of only one fork to replicate regions that could be 100 to 600 kb long. A second model proposes a cascade activation of origins over the TTR[24]. Recent studies suggest that both patterns exist[25].

a. Replication timing and cell differentiation

In mouse, comparison of mouse embryonic stem cells (mESCs) with neural precursor cells (nPCs) derived from the mESCs unravels the modulation of the replication-timing program during cell differentiation. mESCs replication-timing domains range from 200 kb to 2Mb. During cell differentiation into nPCs, some regions initially replicated early become late (EtoL) and similarly some late replicated regions become early (LtoE). These transitions tend to form larger domains of similar replication timing leading to the "consolidation" of replication timing (RT) domains and thus a global increase of RT domains size[26]. These switched zones were ideal to try to identify genomic features coupled to the replication timing definition. As described earlier, the GC richness and gene richness were found associated with timing domains and this tendency is increased upon differentiation. However, the existence of RT domains switches argue against the fact that only the sequence composition would be responsible for the RT program establishment. Interestingly the modified domains exhibit a high GC richness with a low gene density.

Investigations of chromatin marks along domains with different RT show correlations of the transcription-associated marks, H3K4me3 and H3K36me3, with the early RT, independently of cell differentiation state. However, no correlations of the late replicating domains with repressive histone marks such as H3K27me3, H3K9me3 and H4K20me3 were found. The H3K27me3 histone mark (deposited by the polycomb complex) correlated also with the early replicating domains. 87% of the promoters carrying the H3K27me3 mark were found in early replicating domain. Those bivalent promoters were not specifically associated with transition domains[26].

Promoters have been sorted according to their CpG density. Interestingly, in the EtoL regions only promoters with high CpG content were maintained active when switched to late replication whereas other promoters were generally repressed. On the other hand, LtoE domains exhibit a general active transcription of the genes regardless of their promoter CpG association. This CpG high content

promoter genes were also found transcribed in remaining late replicating domains during differentiation, suggesting no clear correlation of transcription of high CpG genes with RT program.

Early replicating domains were found localised in the interior of the nucleus, with euchromatin domains. Cells differentiation transition domains EtoL and LtoE exhibit domains movement toward or away from the nuclear periphery. On the contrary, late replicating domains tend to associate with the nuclear periphery and/or associated with the lamina (Figure 10). Such lamin-associated domains (LADs) were found to be replicated late. The correlation of the LADs with the replication timing, suggested a relation of the RT with the 3D structure of the genome. Hi-C cartography revealed two independent compartments of interaction named A and B. In one compartment, contacts between sequences are enriched but depleted for the sequences of the other compartment. A compartment is depicted as open chromatin in opposition to B compartment enriched with heterochromatin. Interestingly, these compartments correlated well with the RT program with A compartment enriched with early replicating domains, and B compartment, enriched with late replicating domains[27]. A higher resolution of Hi-C data revealed the organisation of genomes into sub-compartments named TADs for topologically associating domains. Comparison of the TADs with the timing replication program find an overlapping of the replication timing domains with the TADs[27,28]. To investigate the implication of TADs on the RT, cohesins and the insulator factor CTCF found at TADS boundaries have been deleted. Although this led to the disappearance of TADs boundaries, no replication timing change could be observed, suggesting that TAD definition is not linked to the replication timing program establishment[29]. However, the organisation into A and B compartments was maintained, probably reflecting a link between this organisation and the RT program.

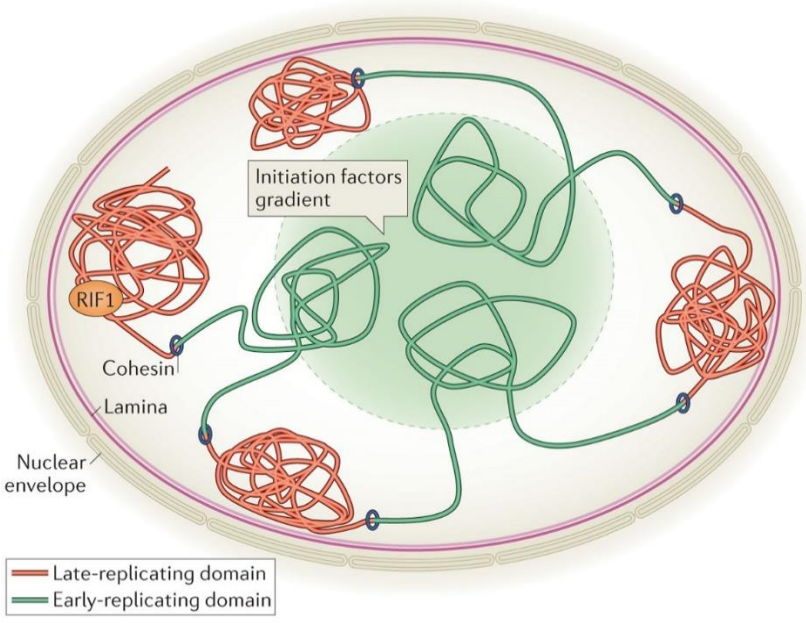


Figure 10

Schematic representation of nuclear organization and replication timing domains. Early replicating regions (in green) are localized at the interior of the nucleus with euchromatin and late replicating domains (in red) are localized at the nuclear periphery close to the lamina and often associated with Rif1.

From Fragkos *et al*, (2015)

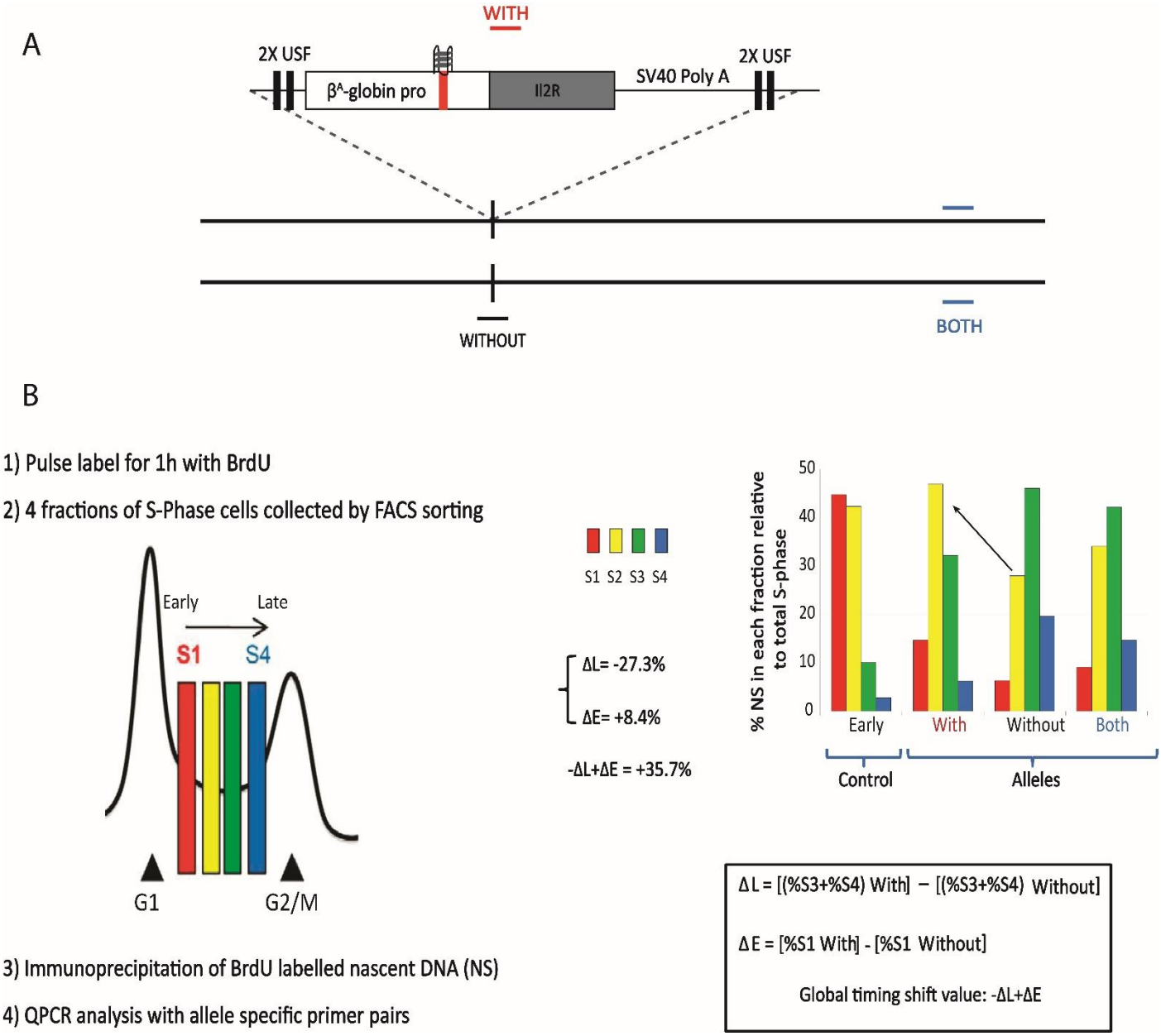


Figure 11

Overview of the replication timing shift assay used to test the capacity of an ectopic origin to advance locally the RT. (A) Schematic representation of the site of insertion of a replication origin construct on one chromosome. (B) Protocol used. On the left, Illustration of fractions selected by FACS according to the cell cycle. On the right, qPCR results from purified newly synthesized DNA.

From Hassan-Zadeh *et al*, (2012)

b. Genetic investigation of replication timing program *cis*-regulatory elements

To identify *cis*-regulatory elements carrying timing information, a construct containing the human β-globin DNAse I hypersensitive sites and the β-globin promoter were fused to the GFP gene and inserted in the Mouse erythroleukemia cells genome. Depending on the construct orientation, the GFP gene could be expressed or silenced, interestingly, when transcribed, the construct could induce a RT advancement but not when silenced. However, the same observation was observed when the β-globin promoter was removed, suggesting mostly the implication of the transcription and DNAse I hypersensitive sites rather than the β^A-globin promoter in this regulation[30].

Another study used the chicken β^A-globin replication origin naturally located in an early domain and ectopically inserted inside a specific mid-late replication region by homologous recombination in the chicken DT40 cell line. The site of insertion is located 300kb upstream of the border of an early domain (Figure 9). The origin, located inside the β^A-globin promoter is fused to the reporter gene IL2R, coupled with the β-actin promoter and the Blasticidine resistance gene[31] (for more detail on the construct see section research project).

The construct's ability to change the insertion site RT was investigated using BrdU incorporation into newly synthesised DNA. After 1h labelling of an asynchronous cell population, cells were sorted into 4 fractions according to their S-phase position, quantification of newly synthesized DNA for each fraction allowed to visualise the construct RT and to compare it to the wild-type allele that conserved its mid-late replication timing. The timing shift amplitudes have been quantified using a calculation comparing the newly synthesized DNA from each fraction from the modified and Wild-type allele (-ΔL+ΔE) (Figure 11).

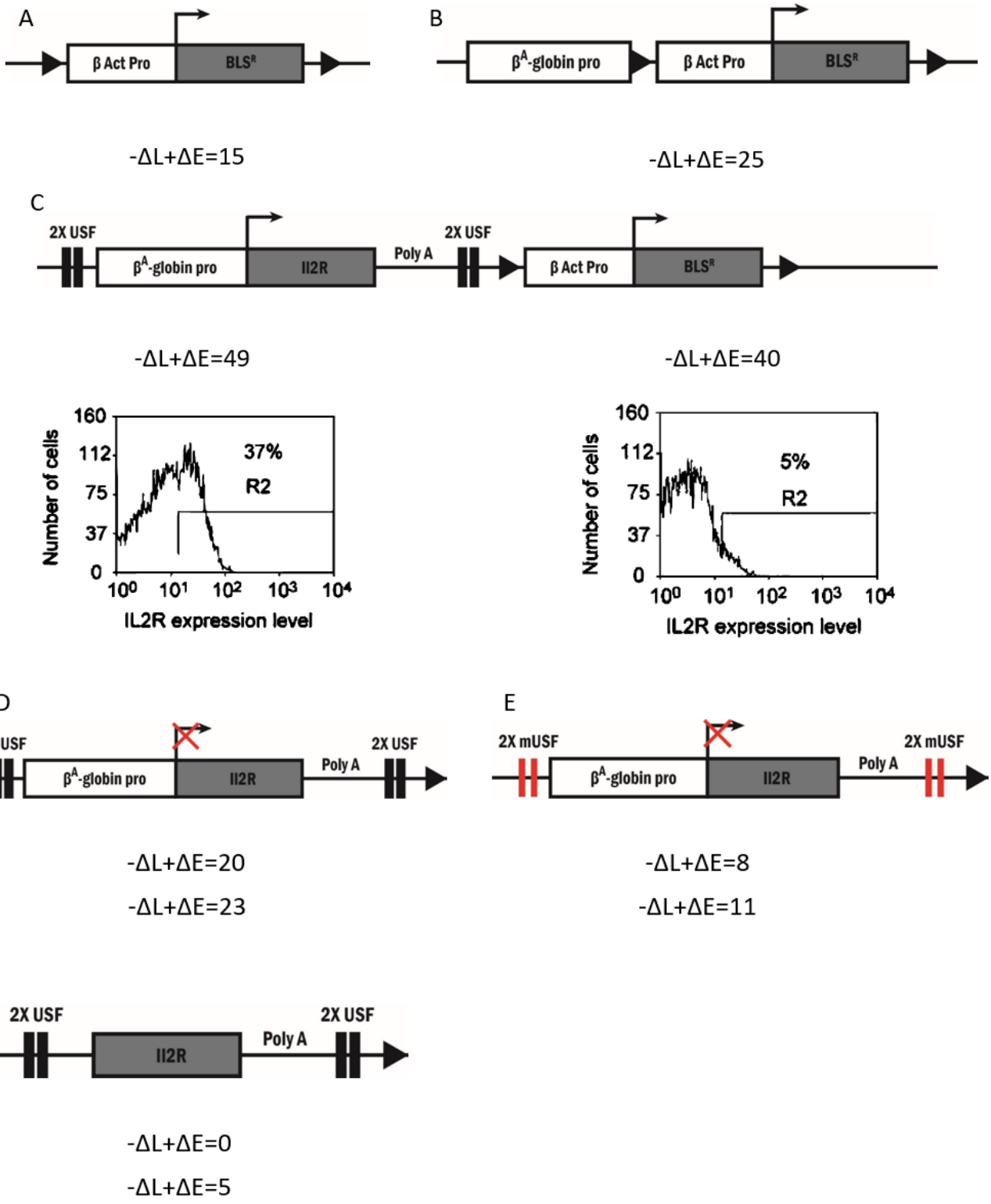


Figure 12

Constructs analyzed in the timing replication shift assay in Hassan-Zadeh et al, 2012. A. Blasticidine resistance gene under β-Actin promoter. B. β^A-globin promoter coupled with BlsR gene and β-Actin promoter. C. Complete construction containing β^A-globin with IL2R gene and SV40 PolyA sequence flanked with 2xUSF binding sites coupled with BlsR gene and β-Actin promoter. Flow cytometry analysis of IL2R protein surface expression D. β^A-globin with IL2R gene and SV40 PolyA sequence flanked with 2xUSF binding sites only. E. β^A-globin construct coupled with mutated USF binding sites. F. IL2R gene flanked with USF binding sites. Arrow head are LoxP sites, -ΔL+ΔE is the quantification of timing shift significant if superior to 15

The blasticidine resistance gene under the control of the β-actin promoter (also containing a strong replication origin) only induces a faint earlier timing shift compared to the WT allele (-ΔL+ΔE=15) (Figure 12A). This number (15) was considered as the limit to define a significant RT shift. Addition of the β^A-globin promoter next to the selection cassette leads to a stronger RT shift (-ΔL+ΔE=25) (Figure 12B). Fusion of the IL2R reporter gene to the β^A-globin promoter reveals that the β^A-globin promoter is slightly active and increases the timing shift when also flanked by USF binding sites (-ΔL+ΔE= 49 and 40) (Figure 12C). **However, after excision of the gene of selection, the β^A-globin promoter with IL2R sequence alone did not exhibit any detectable expression of IL2R gene, suggesting a role of the strong β-actin promoter in this induction.**

The β^A-globin promoter fused to IL2R can only induce a timing shift when flanked by two USF binding sites (-ΔL+ΔE= 20 and 23) (USF being an insulator protein) (Figure 12D). Mutation of USF binding sites abolishes the timing shift (-ΔL+ΔE=8 and 11) (Figure 12E). Finally, as expected the same construct without the β^A-globin origin cannot induce a timing shift as it requires an active replication origin to fire the earlier replication (-ΔL+ΔE=0 and 5) (Figure 12F). In the absence of origin activity, , the construct is passively replicated by the incoming fork from the origins around[31]. **This study unravels cis-regulatory elements critical in the establishment of RT program such as USF binding sites when located nearby efficient origins. It also underlines that transcription of a small gene (~ 1kb) is not sufficient to induce a strong RT shift**[31]. More recently, it was observed that the induction of transcription under the same β-actin promoter significantly advanced the replication timing of a late region when the transcribed gene was very long (~600 kb). This gene size effect suggests that transcription elongation somehow impacts on the RT[32]. Lamina associated gene investigation showed that, once expressed, they loosed their interaction with the lamina and localise inside the nucleus, the relocalisation being correlated with the expression level of the gene[33]. This relocalisation was associated with an earlier replication of the gene. This proximity and state could leads firing factors to bind the expressed gene.

In mESCs, the Dppa2/4 TAD contains three genes Dppa2, Dppa4 and Morc1 that are expressed in mESCs and repressed when cells differentiate. This TAD is replicated early in mESCs and late in somatic cells. In order to identify the *cis*-elements involved in the definition of the timing program large deletions ranging from 2kb to 500kb have been made[34]. As the TAD borders were thought to be involved in the replication program establishment, CTCF sites located at boundaries were deleted. This deletion affects TAD boundaries without changing RT. Similarly, deletion of CTCF protein by a degron does not affect the RT. However, focusing on the intra-TAD contact, three major sites were detected, annotated as containing **Oct4, Sox2 and Nanog binding sites, that are pluripotency major factors. Deletion of these three cis-regulatory elements leads to the delay of the replication timing in mESCs**

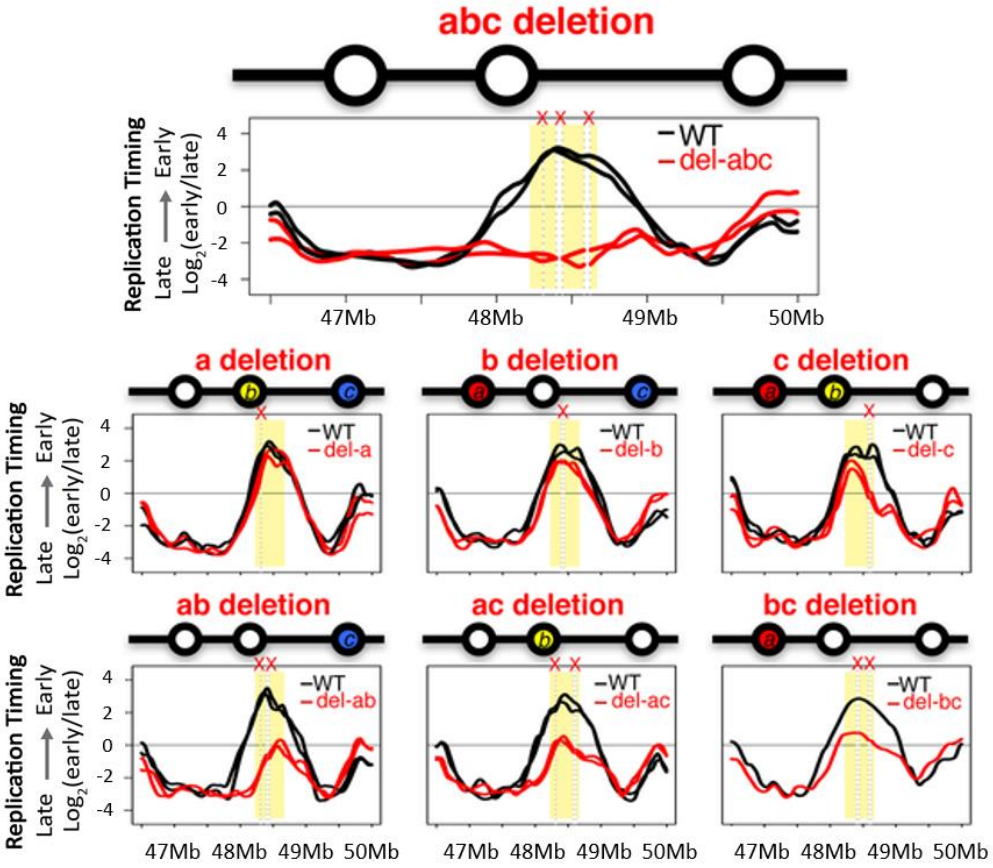


Figure 13

Mouse embryonic stem cells replication timing profiles of chromosome 16 Dppa2/4 TAD locus obtained in WT and mutant cells deleted for a only a, b and c *cis*-regulatory elements or combinations two and three of these elements.

From Sima *et al*, (2019)

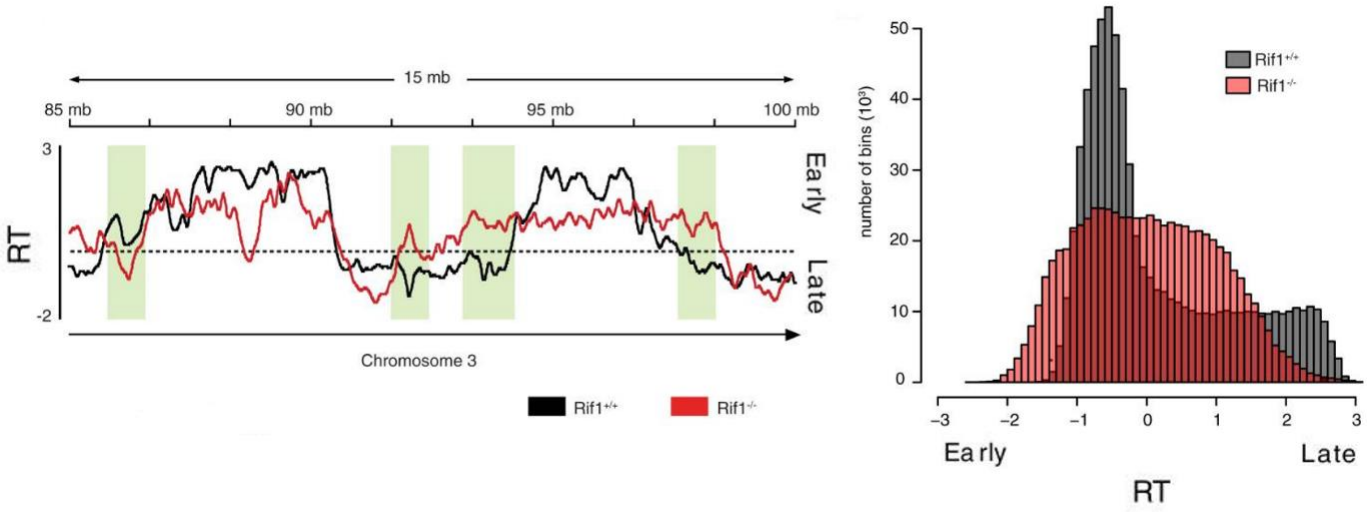


Figure 14

Replication timing profile of mESCs chromosome 3 in a Rif1$^{+/+}$ (black) or Rif1$^{-/-}$ (Red) context. Global replication timing of the genome. A bimodal distribution of the replication timing (early and late) is observed in Rif1$^{+/+}$ cells (grey) whereas an average distribution is found in Rif1$^{-/-}$ cells (red).

From Foti *et al*, (2016)

(Figure 13), nonetheless, TAD boundaries seems to be less well defined. Moreover, deletions of the Nanog, Oct4 and Sox2 sequences affect transcription of the genes found in the TAD, but the relation between transcription and replication timing does not seem to be fully linked as changes in transcription does not always affect the replication timing. To confirm the **Oct4, Sox2 and Nanog role, deletions of similar sequences in other regions also locally affect the replication timing in mESCs**[34]. However, it is important to note that the size of deletions made in this study (30 kb inside the Dppa2/4 TAD) did not allow a clear delineation of critical *cis*-elements.

c. *Trans*-factors involved in replication timing definition

As for *cis*-regulatory elements, *trans*-regulatory factors have been found, their mutation affecting deeply the replication-timing program. I will focus on the most characterised factors: Rif1 and budding yeast Forkhead proteins (Fkh1 and Fkh2).

The Rif1 protein is a well conserved protein found among all eukaryotes and which has been studied extensively in yeasts S. *Cerevisiae* and in S. *Pombe*. Rif1 deletion drastically affects the RT program with late origins replicated earlier, but also early origins replicated later. **Rif1 deletion induces the replication timing profile averaging, with less clear early and late domains** (Figure 14). Rif1 was firstly found associated with telomeres in yeasts. However, Rif1 can also bind DNA at intergenic regions[24]. 80 Rif1 binding sites were detected not necessarily overlapping with replication origins, suggesting that Rif1 could repress several replication origins at the same time[24]. It has been shown in fission yeast *S. pombe* that Rif1 interacts with the Protein Phosphatase 1 (PP1) that dephosphorylates DDK-dependent phosphorylation of the Pre-RC MCM subunits resulting in the delay of the firing until late S-phase[35]. Rif1 C-terminal can also interact with Dbf4 essential protein for the DDK activity, interaction that is dispensable for replication repression. Interestingly it appears that Rif1-PP1 interaction was disrupted by Rif1 DDK phosphorylation[36]. Rif1-PP1 interaction appears to be even more complex has Rif1-PP1 deletions decrease the origin licensing in G1[35].

Rif1 protein is involved in several other mechanisms such as telomere maintenance and non-homologous end joining repair. In mESCs, Rif1 is found enriched at late-replicating domains, creating Rif1-associated domains (RADs). These domains overlap extensively with Lamina associated domains (LADs) representing 73% of total late replicating domains[37]. Among those late replicating domains, two subtypes can be distinguished, the Rif1 and Lamina associated domains and the only Rif1 associated domains. Interestingly, Rif1 deletion affects mostly the RADs that are not overlapping with LADs. The RADs and LADs late replicating domains maintain their RT upon Rif1 deletion, suggesting that Rif1 is not required or is simply redundant with a second layer of RT regulation associated with nuclear lamina organisation[37]. Rif1 deletion leads, after a long-term proliferation, to a mis-regulation of transcription

of some genes. However, those genes were not found close to Rif1 binding sites. Moreover, the epigenetic landscape of affected genes were not affected even in replication timing program disorganisation context.

Rif1 mutation, however, exhibits an altered genome 3D organisation with the establishment of new contacts from the RADs with domains that were not contacted in wild-type context. To control that these new domains interactions were not due to the replication program modification, such approach was used before a first round of replication[37]. Interestingly, new interaction between TADs could be observed before any replication, **which would suggest that in addition of replication timing program regulation, Rif1 also play a role in genome 3D organisation**.

The large and well conserved family of Forkhead (Fkh) transcription factors play a major role in several mechanisms. Interestingly, in budding yeast **Fkh1 and Fkh2 deletions were found to delay the firing of a subset of early replication origins whereas, some late replication origins fired earlier**[38]. From 352 fired origins in budding yeast, 106 (30%) had their origin activity decreased (Fkh-activated) and 82 (23%) had their origin activity increased (Fkh-repressed). Such phenotype was observed when both Fkh1/2 were deleted. Deletion of Fkh1 affected only a subset of origins, whereas Fkh2 deletion did not have any detectable effect. Fkh1/2 deletions, delayed some early origins, but interestingly, another subsets of origins replicated earlier. Deep investigation of the chromosomes organisation uncovered the clustering of Fkh-activated origins with Cdc45 (firing factor). Fkh-binding sites were found enriched at Fkh-activated origins and depleted at Fkh-repressed origins. However, Fkh-binding sites were also found enriched at unaffected origins. Investigation of the Fkh-binding sites repartition over the origins showed a close proximity of Fkh-binding sites with the ACS in Fkh-activated origins compared to unaffected origins that exhibit more distant Fkh-binding sites[38]. Mutations in the Fkh-binding sites altered the Fkh fixation without affecting ORC binding and changed the RT firing of the nearby origins. Replication timing program modifications were unrelated to modified gene transcription levels around the origins.

The mechanism by which Fkh1/2 induced early origin firing has been investigated molecularly. Fkh1/2 can recruit Dbf4 through direct protein-protein interactions and thus leads to origin firing nearby through DDK activity. Dbf4 interaction with Fkh1 was dependent on Fkh1 C-terminus part, deletion of this region mimics the Fkh1 deletion phenotype, indicating that Dbf4 recruitment is necessary to early replicating origin but not essential for the replication of the rest of the genome. Dbf4 fusion with Fkh DNA recognition motif rescues the Fkh1/2 deletion phenotype. Dbf4 is also responsible for the CDC45 factor recruitment to the firing Pre-RCs[39]. CDC45 ChIP reveals its localisation at pre-RCs during G1 phase even though its recruitment to Pre-RCs is DDK-dependent (DDK not being express in G1). Such

CDC45 recruitment presages the early firing of these origins. Among the mapped CDC45, we find the Fkh-activated origin[38]. Fkh1/2 deletion leads to the loss of CDC45 enrichment at the Fkh-activated origin, which could be due to the loss of Dbf4 recruitment.

Fkh factors associated with replication origins create clusters of origins favouring the early firing and the factors recruitment[38]. Early replicating origins seem to be relocalised in a Fkh-dependent manner (through DDK activity) in the nucleus in G1 phase from the nucleus periphery to the nucleus interior[40].

Taken together, these results clearly show a very complex regulation of replication timing even in the yeast S.*Cerevisiae*. Further studies will be needed to understand all the potential mechanisms involved in the establishment of the highly precise temporal program of DNA replication. Another important unanswered question remains, which is, what is the role of this temporal program. Is it crucial to maintain genomic integrity and/or epigenetic memory?

3. Replication origin studies

One important step to study replication origins was the mapping genome-wide of their localisation. To answer this question, several technics came out, starting from plasmid transmission from mother cells to daughter cells to long DNA molecule sequencing using Pacific Bioscience technology, passing through bubble trapping, Short-Nascent Strand purification and ORC/MCM Chromatin Immunoprecipitations. All of these technics relied on different replication initiation aspects and brought a number of consistent elements to identify replication origins. Here I will describe a non-exhaustive number of technics and compare them.

a. ORC / MCM Chromatin immunoprecipitation

One strategy to identify replication origins in metazoan was to detect Pre-RCs sub-units. The ChIP approach should identify their positions on genomic DNA and therefore the sequences and/or chromatin elements required for Pre-RC binding.

The first ChIP experiment made in human was on the ORC1 sub-unit. In this study, the authors identified nearly 13 000 peaks from Hela cells, where eleven extracted regions were enriched in small nascent strands (SNS) and co-localize with ORC2 and MCM5. However, in this study, input DNA has been submitted to a gradient purification before the ChIP to recover only the “Low-density DNA” corresponding mostly to more open chromatin. Therefore this analysis most probably excludes a part of ORC1 recognition sites found in more condensed regions[41].

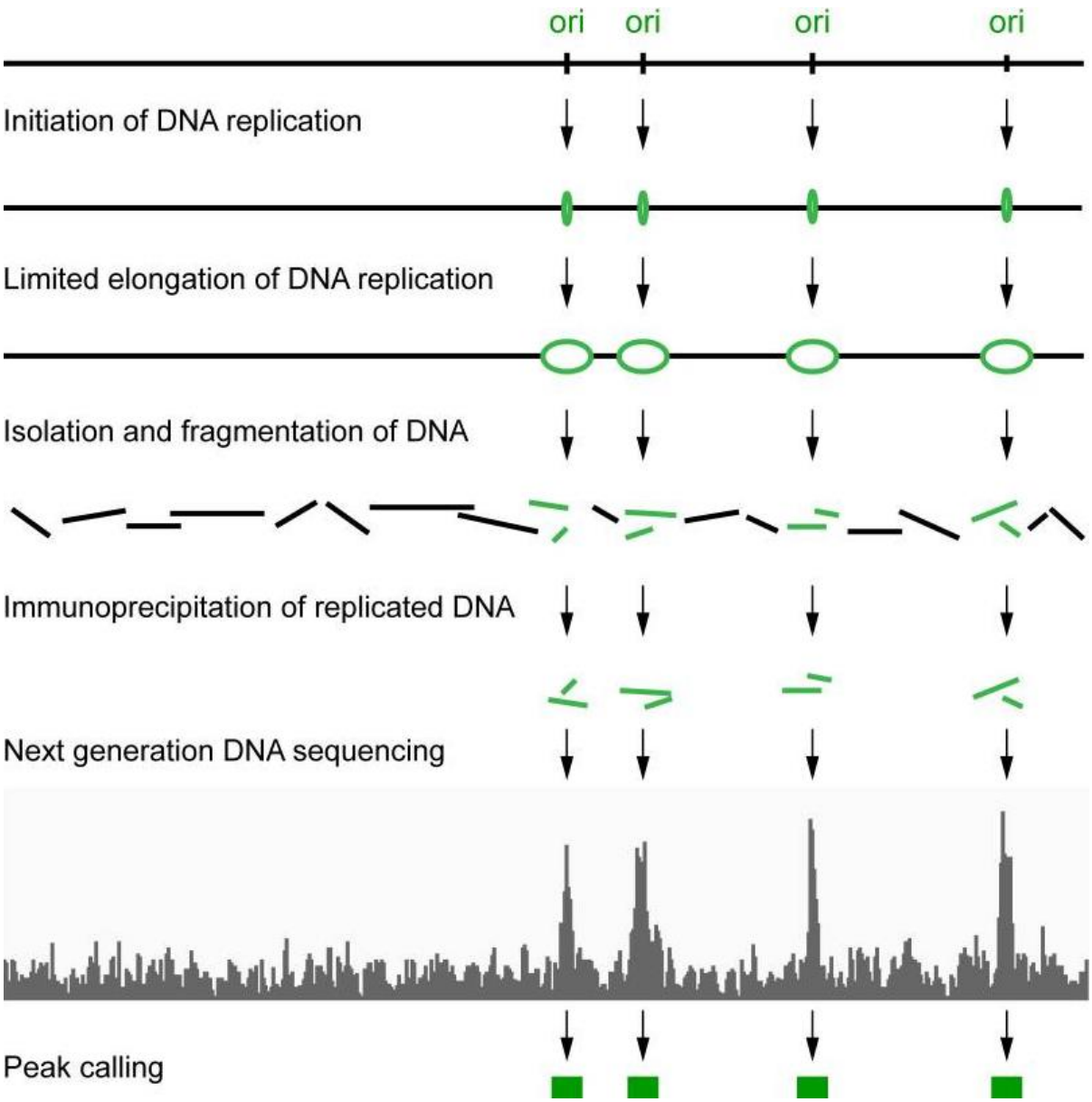


Figure 15

Overview of the Ini-seq experimental procedure, green fragments represent labelled DNA by digoxigenin-dUTP.

From Langley *et al*, (2016)

ORC2 ChIP has been made in K562 human cells, this time on global genomic DNA. Around 52 000 peaks were identified[42]. Cross-validation of this new ORC2 ChIP data set with ORC1 ChIP data was made and shows a good overlap, cross-validating each experiment. However, ORC1 ChIP, as discussed before, was submitted to a first round of DNA fragmentation and purification, and was made in a different cell type. In order to compare the two datasets, the authors used histone marks associated to ORC1 binding sites in Hela cells and compare this data with an ORC2-based machine learning program scoring ORC2 potential binding sites. No recovery percentage nor details on the algorithm selection windows were indicated. From the algorithm proposed during the study, it came out that the combination of open chromatin regions (sensitive to DNAse I hypersensitivity), histone H3 acetylation and methylated H3K4 were predictive of ORC binding[42].

Pre-RC is constituted of the ORC complex but also of the MCM helicase, the last one has been used to ChIP Pre-RC position in HeLa cells. The sub-unit MCM7 has been used for ChIP-seq experiments and identified nearly 200 000 peaks present in two biological replicates[43].

Interestingly, a quite important discrepancy could be observed between ORC and MCM peaks localisations in the different experiments, even though, the proteins used are part of the Pre-RC complex. These differences sign a very dynamic process during the origin licensing and firing. It has been reported that ORC could deposit several double hexamer helicases[44] and that helicases can be displaced by gene transcription[44,45], elements that could explain the discrepancies observed in those data sets.

b. Initiation sequencing

Also called Ini-seq, this cell-free system, based on digoxigenin-dUTP bases incorporation into newly synthesized DNA identified around 25 000 peaks [46]. Bladder human cells EJ30 were subjected to G1/S synchronisation using mimosine. Extracted nuclei were then mixed with proliferating human cells extract, containing essential DNA replication proteins, coupled with desoxy-nucleoside triphosphate (dNTP) and digoxigenine labelled dUTP. This cell-free system can support DNA replication at already licensed replication origins. DNA replication allowed modified bases incorporation into neo-synthesized DNA fragments. After a short pulse, labelled DNA was purified using an anti-digoxigenine antibody and then sequenced (Figure 15).

One limitation of the Ini-seq method is that origins identified are only early firing ones. Accordingly, peaks repartition direct analysis among the different timing region illustrates a predominantly cluster in early replicating region (75-80%) compared to late replicating region (10-15%)[46]. Such technic provides a population-based map of replication origins. Moreover, Ini-seq cartography gives a discrete origin map and origin strength information derived from the peaks height.

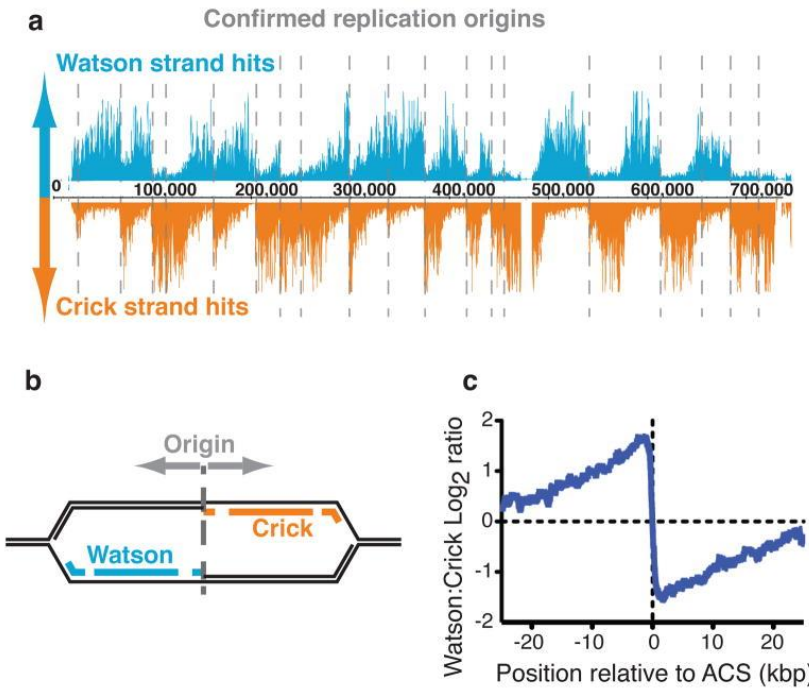


Figure 16

Typical Okazaki-seq outcome in budding yeast. Origins are identified by the observed drastic skew from Watson to Crick strands hits. Vertical dot grey line are Watson to Crick skew unraveling a replication origin.

From smith *et al*, (2012)

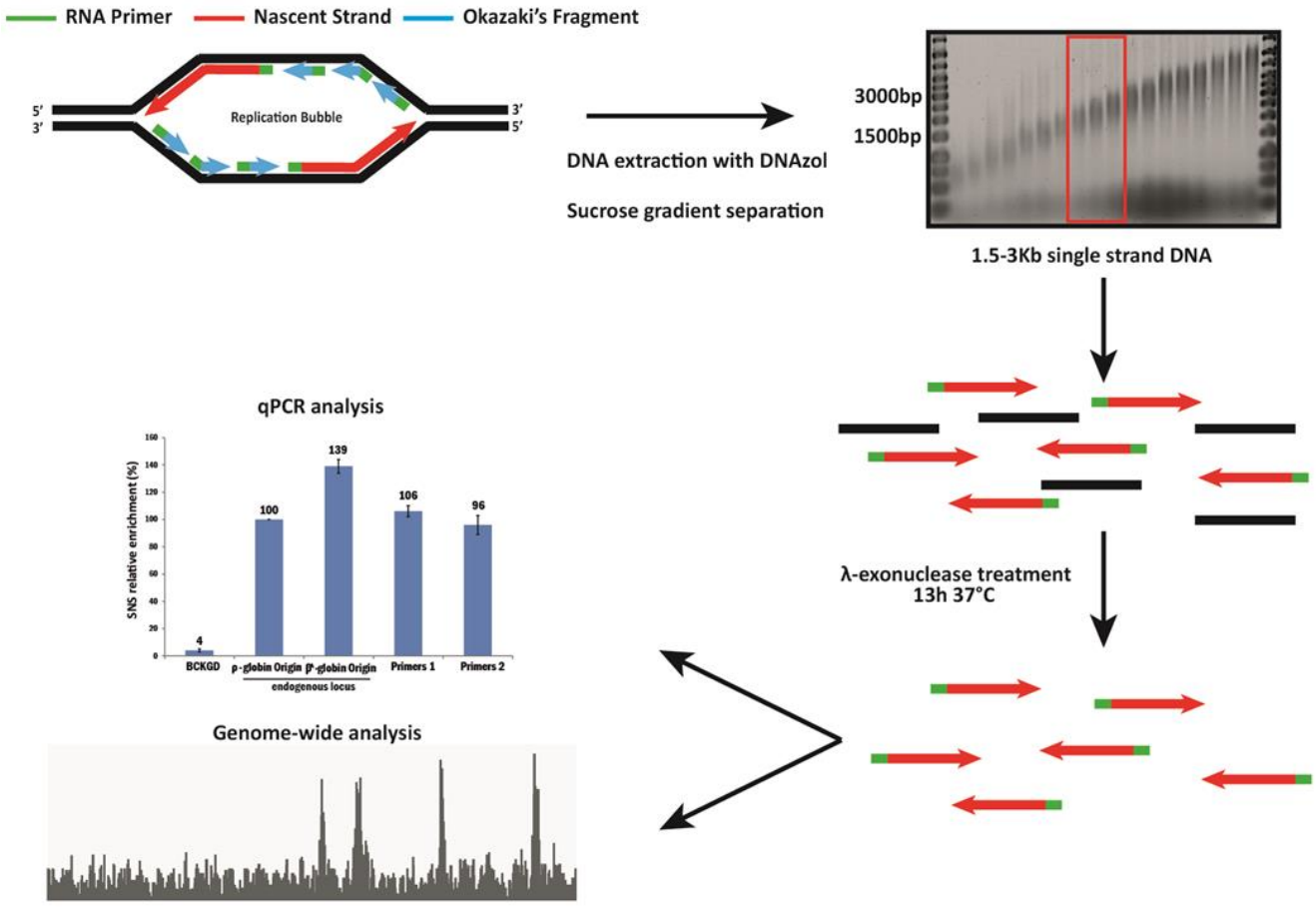


Figure 17

Overview of the SNS purification technic. Nascent strands are shown in red with their RNA primers in green. After DNA extraction, ssDNA is subjected to a sucrose gradient and 1.5 to 3 kb long fragments are recovered. These fragments are a mixture of SNS and genomic broken DNA. After λ-exonuclease digestion, SNS are sequenced or quantified by qPCR.

c. Okazaki fragments sequencing

Developed in S. *Cerevisiae*, this technic relies on the purification of Okazaki fragments that emerged from the replication fork on the lagging strand. In S. *Cerevisiae* Okazaki fragments purification relies on the DNA ligase I deletion that results in the lack of ligation between Okazaki fragments [47]. Okazaki fragments were then purified and sequenced. In human cells in order to avoid DNA ligase I deletion, the authors took advantage of short-pulse DNA labelling by modified DNA-bases incorporation during DNA replication. Once labelled DNA was extracted, Okazaki fragments were purified depending on their size (around 200bp) and immunoprecipitated with specific antibody directed against modified bases. Immuniprecipitated DNA was then sequenced in a strand specific manner. Replication origins are characterised by a Watson to Crick transition of OK-seq profiles, also giving at an unprecedented level of resolution the fork orientation[48,49]. Okazaki fragments mapping permits to identify initiation sites with clear Watson to Crick transition and initiation zones where Watson to Crick transition was not contained in a small window, but changes in a broader way, resulting of a multi origin firing area. A Crick to Watson transition also defines replication termination zones[50]. (Figure 16).

OK-seq provides a population based replication origin map, using asynchronous cell population. Okazaki fragments are obtained with a quite high level of background resulting from the progressing fork and not only from firing origin. Compared to the yeast S. *Cerevisiae*, where origins are well defined and isolated, in metazoan, such feature is rare and does not allow the identification of many isolated origins.

d. Short-Nascent Strand purification

Based on replication-specific DNA intermediates molecules, this technic relies on the purification and relative quantification of short nascent leading strands purification emerging from fired replication origins[51]. Short Nascent Strands (SNS) possess, such as Okazaki fragments, at their 5' extremities, a RNA primer deposited by the primase and used by DNA polymerases to initiate DNA replication. From an asynchronous cell population, genomic DNA, containing replication bubble, is gently extracted, heat denatured and submitted to size selection on a sucrose gradient. Molecules of around 1-2.5Kb are recovered to avoid Okazaki fragments recovery (around 200bp) and long molecules coming from potential nearby replication origins (> 5kb). However, recovered DNA is a mixture of SNS but also of contaminating broken genomic DNA. The Use of λ-exonuclease allows the digestion of un-RNA primed DNA, resulting in the hydrolysis of un-relevant genomic DNA (Figure 17).

SNS relative enrichment was the first technic used to provide a human replication origin map at a genome-scale level. SNS can identify discrete replication origin peaks but requires strong origin efficiency to ensure their detections[52–54]. Moreover, concerns were raised about the λ-exonuclease

activity and specificity during SNS purification. Indeed, λ-exonuclease appears to have a bias due to its potential difficulty to digest GC rich sequences and more specifically DNA secondary structures such as G-quadruplex[55]. However, an appropriate use of high λ-exonuclease concentrations during SNS purification overcomes this potential bias.

e. Summarize of the different origin replication study technics

The technics described earlier brought to the field of DNA replication an unprecedented level of information on DNA replication origins. The high spatial resolution of some methods permits the identification of several features associated with replication origins. However, the discrepancies between the different data sets have raised some concerns about the bona fide identification of replication origin in the different technics used.

SNSs data comparison with ORC ChIP data shows a modest correlation with only 13% of SNS enrichment within 1kb windows of ORC2 binding sites and 41% of SNS peaks located within 10 kb[42]. The ORC1 data set, even though it is not a complete landscape of ORC1 binding sites, has been found to overlap with 44% of replication initiation sites detected by SNS[56]. Interestingly, a much higher correlation between MCM peaks and SNS data set was observed with 86% of SNS peaks found associated with MCM peaks (only 39% of MCM peaks found enriched at SNS peaks) (Figure 18). **The correlation of MCM Peaks at SNS peaks level can be partially explained by the direct contribution of MCM to open the DNA double helix at replication origins, moving away from the replication origin**. Moreover, from an excess of loaded pre-RCs during G1 phase, only a subset are fired according to the so called Jesuit model "Many are called, few are chosen"[57]. This hypothesis explains that only 39% of MCM peaks correlate with SNS mapped origins, the remaining non-SNS associated MCM peaks were therefore considered as dormant origins[43]. Dormant origins are back-up origins that fire in case of DNA replication stalling to rescue the DNA replication. **MCM dissociation and mobility from ORC could be the reason for the weak ORC-SNS peaks correlation.** Other possibilities are that ORC has another function such as heterochromatin formation or/and that ORC does not remain strongly associate with the MCM double hexamer after its deposition.

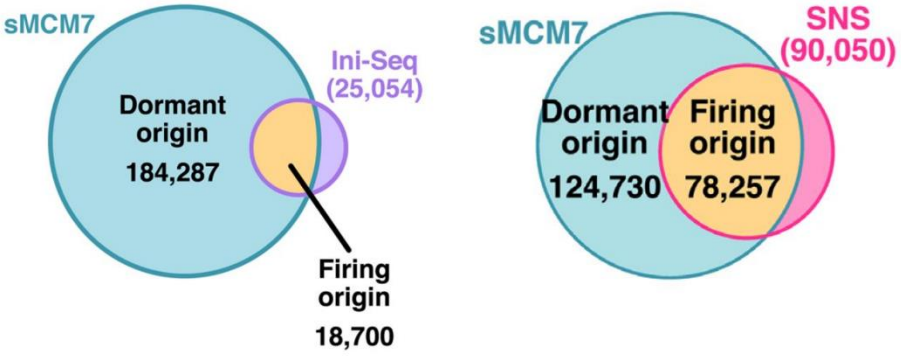


Figure 18

Overlapping of SNS-seq and Ini-Seq peaks with MCM ChIP-seq results.

From Sugimoto *et al*, (2017)

In the first published OK-Seq paper, a strong emphasis was put on OK-seq and SNS-Seq discrepancies with only few SNS recovery within the initiation zones[48]. However, in a second OK-seq analysis[50] a strong enrichment of replication origins at promoters, where SNS peaks were also found localised suggested more agreement between the two methods. Observed discrepancies can be explained by the different degree of sensitivity and resolution of each method (SNS and OK-seq) coupled with a different way to filter data sets during genome-wide comparisons of the two methods. The important result that emerges from these studies is that both methods could detect promoters and more specifically actively transcribed promoters as being strong sites of replication initiation.

Ini-seq data, obtained without the use of λ-exonuclease, show a good correlation with SNS-seq, with 56% of Ini-seq sites overlapping with SNS peaks when two different cell lines are compared (only 14% of SNS peaks correlates with Ini-seq) (Figure 19). The differences in data sets overlapping can be explained by the lack of detection of mid and late replicating origins in Ini-seq experiments. Indeed, in SNS-seq, molecules are extracted from an asynchronous cell population allowing detection of strong origins from early to late-S. By contrast, in the ini-seq protocol, G1 cell synchronisation allows only detection of mostly early firing origins. Surprisingly, Ini-seq data overlap with OK-seq initiation zones to 36% and 22% (depending on the Ini-seq replicates) (Figure 19), suggesting the presence of discrete replication origins in those initiation zones.

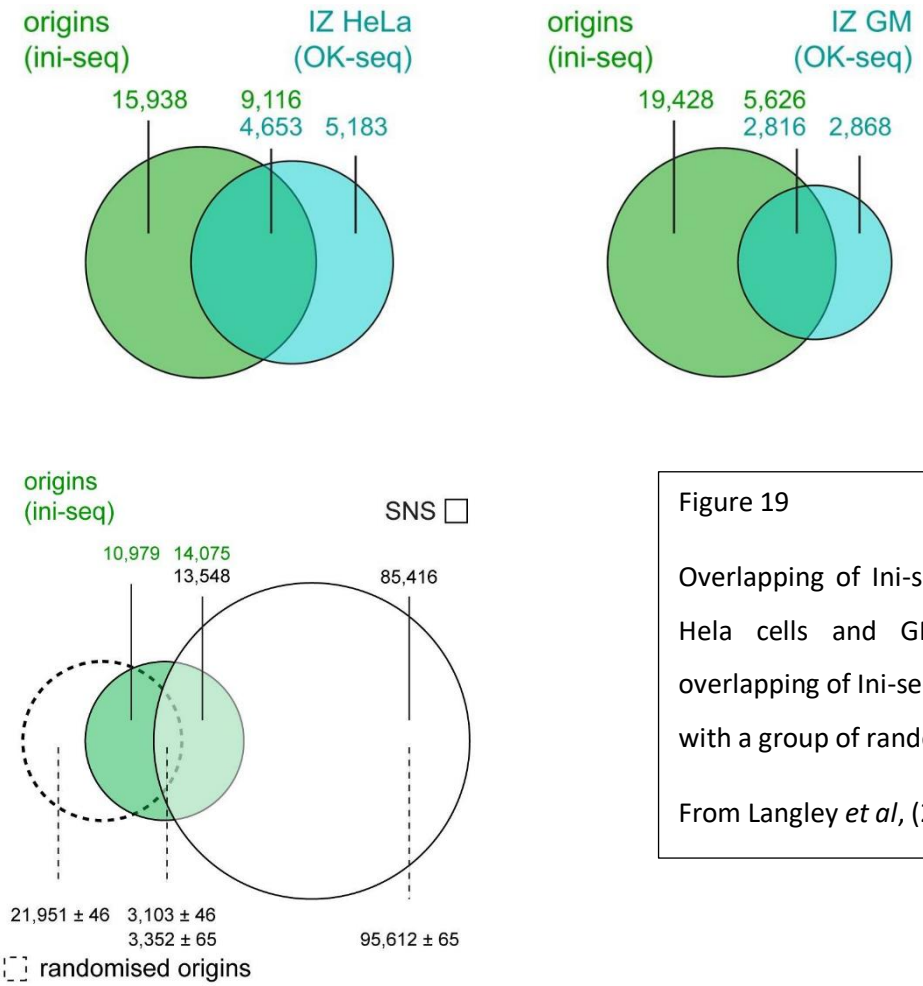


Figure 19

Overlapping of Ini-seq and OK-Seq peaks from Hela cells and GM06990. At the bottom, overlapping of Ini-seq with SNS-seq peaks shown with a group of randomized Ini-seq origins.

From Langley *et al*, (2016)

4. Replication origins in eukaryotes

Since the depiction of the replicon model in 1963, great effort has been made to uncover replication origins sequences, through genomic to genetic studies. Although a two-step replication origin loading process and a great conservation of replication proteins are shared between yeasts and vertebrates, replication origins are not conserved from the yeast *S. cerevisiae* to human and show very distinct features.

a. Characteristics of Autonomous Replication Sequences in *S. cerevisiae*

Identified in 1979 by a plasmid transmission experiment, the first autonomous replicating sequence (ARS) identified was a 1.4 kb long sequence, precise deletions decreased its size to 850 bp[58]. Extensive study of other ARS[59] allowed the identification of an 11bp consensus sequence entitle autonomous replicating consensus sequence (ACS) that will be later on called sequence A. ACS is essential but does not support DNA replication by itself, it requires a sequence B to ensure DNA replication (Figure 20). A third sequence, called C, has also been observed, but its role relies on the B sequence presence[60]. The B domain has been then divided into three sub-sequences B1, B2 and B3 and is located 3' of the ACS. The sequence C is not found at all ARS, as the B2 and B3 sequences, but the C sequence is usually found 5' of the ACS. Sequences A and B1 turned out to be recognised by the origin recognition complex that is essential for DNA replication[61], sequence B2, present in most, but not all ARSs, is used to load the Mini-chromosome maintenance protein, both proteins are essential proteins of DNA replication. B3 acts as a DNA replication enhancer bound by the transcription factor called ARS-binding factor 1 (ABF1). High throughput analyses of DNA replication in yeast identified around 12 000 potential ACS, however, only around 400 ACS[62] are activated during the cell cycle. It has been observed that ARS exhibit an A/T skew, with a strong T enrichment 5' of the ACS and a shift to A enriched at the 3' part of the ACS[63], this A/T skew is observed in a limited zone of 100 bp. In a much broader area, a quite high GC-skew could be observed that went back to zero in a 1000 bp window[64]. No clear histone tail modification was observed at the ARS when considering all the origins from early to late.

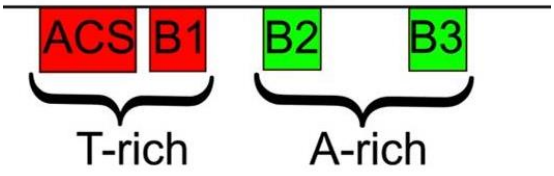


Figure 20

ARS1 structure

From Eaton *et al*, (2010)

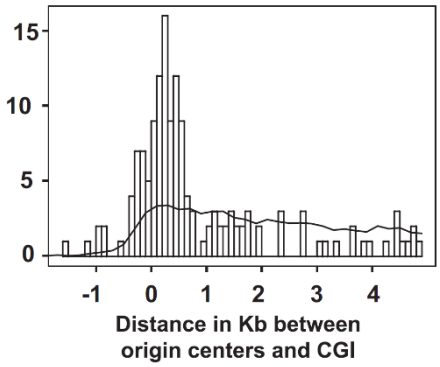


Figure 21

SNS identified origin distribution around CGI in human.

From Cadoret *et al*, (2008)

Mechanistic study of ARS *in vivo* and *in vitro* investigated the Pre-RC loading dynamics as well as the origin activity according to ARS orientation and structure. Interestingly, the A and B1 sequences, lacking the B2 as well as the enhancer B3 and C elements, cannot sustain replication activity (and MCM loading). The Pre-RC loading and origin activity were restored only with two A and B1 sequences in a face-to-face orientation. Replication activity is rescued by the presence of a second A and B1 sequence positioned in a head-to-head manner, other orientation could not rescue the replication activity. Similarly, distance between the two genetic elements from 25 bp to 400 bp exhibit a replication origin activity with a slow decreasing efficiency from 25 to 400 bp. Most efficient MCM loading peaks could be observed with a 70 bp distance between the two head-to-head ARS. Signing a Pre-RC improved loading, that however, does not affect the replication origin activity drastically[65].

b. Genomic characterisation of replication origins in metazoan

i. Genomic features associated with replication origins

Purification of SNS coupled with deep sequencing allowed the identification at an un-precedent level of a high number of replication origins. Deep characterisation of sequences associated with SNS peaks failed to identify a consensus motif as in the yeast *S. cerevisiae*. The first large scale study of SNS peaks has been made in Hela cells and allowed the identification of 282 origins along 1% of the human genome (ENCODE regions)[52]. Replication origins correlated with GC-rich sequences and were strongly associated with CpG Island (CGI) (Figure 21). CpGs are important epigenetic features among metazoan; cytosine methylation affects the chromatin and gene expression. CGI are defined by a sequence of at least 200 bp with 60% of CG and having a CpG observed / CpG expected superior to 0.6[54]. Among the 282 identified origins, 1/3rd were overlapping CGIs and half of the origins were located 1kb away from a CGI. Another origin important identified feature was their association with promoters. Indeed, 34% of identified origins were associated with promoter, among which 79% were associated with CGI. A similar study in mouse Embryonic Stem Cells (mES) identified 97 origins, as in human, replication origins were found associated with transcriptional units (85%) and more precisely to promoter (44%) from which 88% correspond to CGI associated promoter[53] (Figure 22). A particular attention has been drawn to origin efficiency and it appeared that origins associated with CGI and promoters (most efficient promoters) were the most efficient ones.

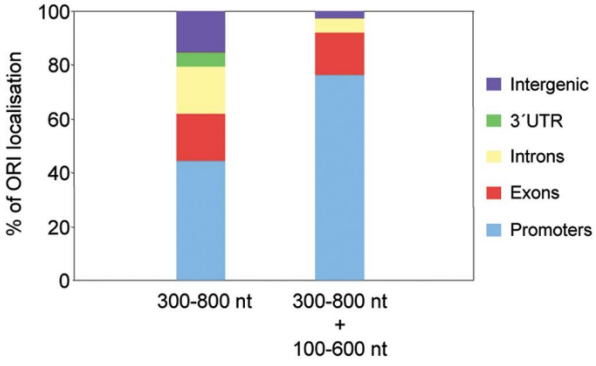


Figure 22

SNS identified origin distribution at genomic elements depending on SNS size selection.

From Sequeira-Mendes *et al*, (2009)

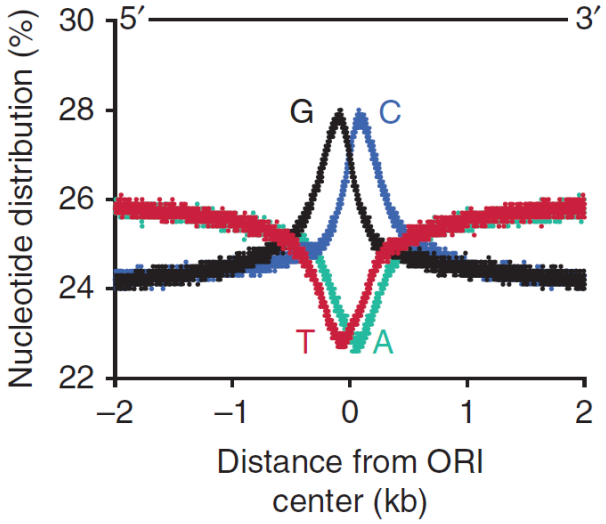


Figure 23

Nucleotide distribution around SNS peaks.

From Besnard *et al*, (2012)

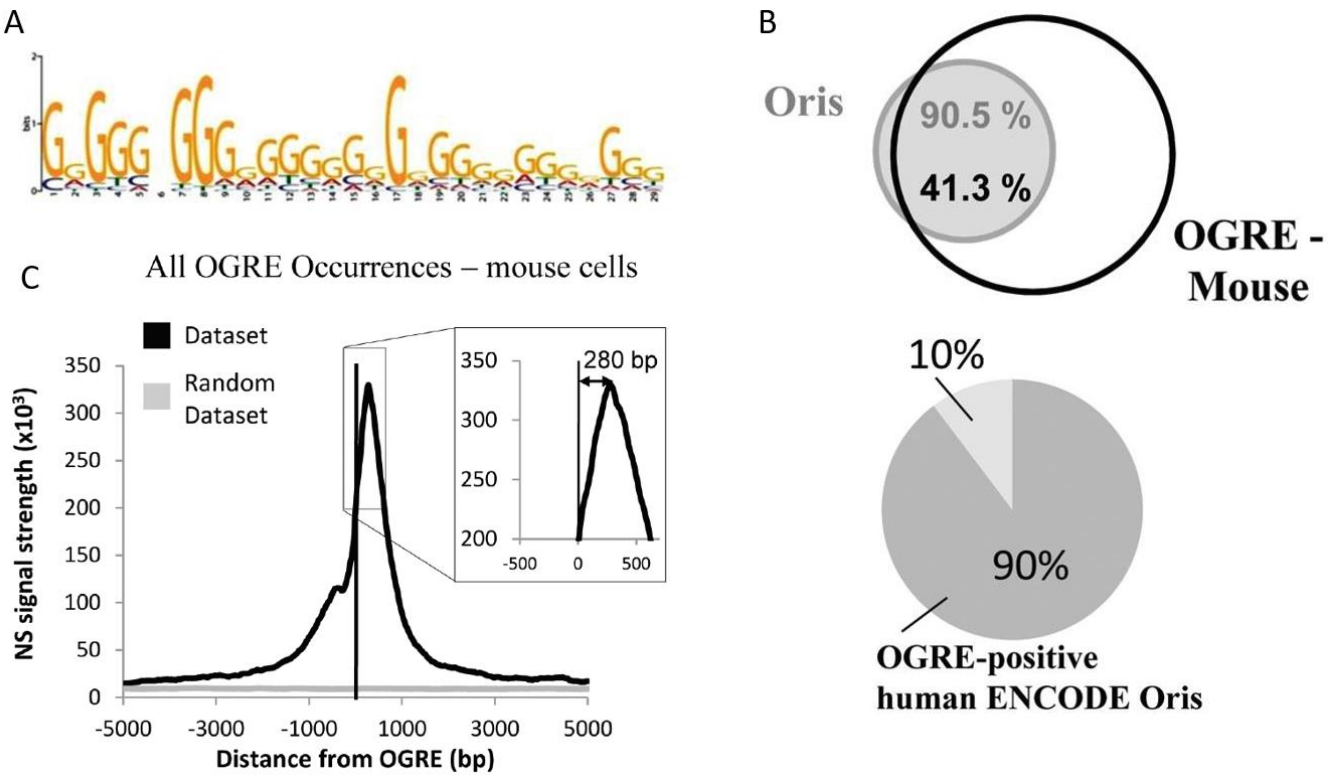


Figure 24

A. Mouse OGRE motif. B. Association of origins with the OGRE sequence (upper) and human origins association with mouse OGRE sequences (Lower). C. SNS enrichment centered on OGRE sequences.

From Cayrou *et al*, (2012)

Those two studies revealed for the first time a predominant class of cis-elements associated with replication origins and confirmed the lack of a consensus sequence for metazoan origins. However, the small number of replication origins, in part due to technical limitations of micro-arrays, did not provide a clear and complete view of replication origins landscapes, although different types of genomic landscapes were inspected. In 2011, a study using three mouse cell lines and one drosophila cell line, identified 13 000 SNS peaks[54]. The increased number of SNS peaks confirmed features that had been already observed in previous studies (27% of origins overlap with a CGI and 64% of CGI are within an origin in mouse). The Drosophila genome does not exhibit CGI, but has similar genomic elements that, as in mouse, show an association with origins (59% of origins associating with CGI-like, and 46% of CGI-like associating with replication origins). However, in Drosophila, replication origins do not associate with TSS[66]. A conclusion of the different observations was that **CGIs outside TSS are associated with origins as well as TSS without CGIs**.

Identification of replication origins in several human cell lines allowed to classify replication origins depending on their usage. Constitutive origins exhibit a similar efficiency independently of the cell line analysed, specific origins, on the contrary, have an efficiency depending on cell type and common origins are defined has origins found in different cell lines but that are not constitutive[67]. Replication origins usage depending on the cell line clearly indicates that replication origins efficiency and activation are dynamic processes that must rely on different activating and repressing elements.

We learned from *S. cerevisiae* that origins are characterised by an A/T skew. Similar skew in metazoan origins has been investigated in mouse and in human and found to be a GC and AT-skew that invert exactly at the SNS peak position. Gs and As were found enriched at the SNS peaks 5' part and Cs and Ts at the peaks 3' part [54,67] (Figure 23). More importantly, this skew led to the discovery of G-rich motifs that have been called OGRE for Origin G rich repeated elements (Figure 24A). Such G-rich motifs were found to potentially be structured into G-quadruplex (pG4)[56,67]. Folding of pG4 occurs through the interaction of four tracts of three Gs, separated by loops, this secondary structure has been initially defined by the consensus sequence (G_3-N_{1-7}-G_3-N_{1-7}-G_3-N_{1-7}-G_3). The structure and G4's characteristics will be discussed more extensively in the next part. In mouse, it appears that 80-90% of origins associate with OGRE motif. The OGRE motifs seems to be a strong replication origin predictor. However, even though origins associate strongly with OGRE motifs, only 41% of all OGRE motifs associate with replication origins. Coupling of OGRE motifs with CGI increases the replication origins association to 90% (Figure 24B). In summary, **replication origins are highly associated with OGRE motifs, but on contrary, OGRE motifs are only partially associated with replication origins**. However, such motifs have been found to be strongly associated with replication origins in human, 67% of origins were found associated with strict consensus G4 motifs. Increasing of loops size from seven to fifteen

increases the origins enrichment with pG4 to 91%. More interestingly, in human, most of the replication origins associated with TSS and CGI were associated with pG4[67]. Analysis of the pG4 density coupled with the origin efficiency underlines a link between the two parameters, indeed, an increasing number of pG4 is associated with an increase in replication origin efficiency[56].

With the discovery of pG4 implication on replication origin, the previously described GC skew parameter has been more extensively studied, addressing the role of pG4s on the skew. To do so, origins have been oriented according to the OGRE sequence leading to a loss of the Cs enrichment 3' of the SNS peaks and thus revealing the direct role of the OGRE sequence on the observed GC skew. Moreover, it appears that OGRE sequences are located on average 280bp upstream of the SNS peaks in mouse (Figure 24C), and 160bp upstream of the SNS peaks in drosophila[68]. This result indicates that although the **OGRE sequence is strongly associated with SNS peaks, it does not seem to be located at the replication initiation starting point**. A similar analysis, combining three model organisms (Mice, Human and Chicken), this time focusing on pG4 rather than OGRE sequence, shows that pG4 localised at 150 bp away from SNS peaks in human and 50 bp in mouse and chicken[69].

A close examination of replication origins on mouse chromosome 11 showed a non-homogenous replication origins repartitions, with regions of high density and regions of low density. This density correlated with the replication-timing program, where the early replicated regions were dense in origin and the late replicating regions exhibited a low replication origin density[54]. Such observation has also been made in human[67]. During cells differentiation, RT program can be modified and early replicating regions can switch to late replicating regions, such timing changing regions showed a less replication origins density profiles compared to un-affected differentiation timing regions[47]. The opposite observation was true for Late to Early replication region.

ii. Epigenetics marks associated with replication origins

After the identification of origin associated *cis*-regulating sequences, one major focus was on the identification of epigenetics mark associated with replication origins.

One study of epigenetics features identification relies on the replication origins sorting according to the number of SNS peaks found in a 14 kb window. This classification defined three major classes in mESC[70]. Class 1 represents isolated SNS peaks and are mostly replicated late (75%), not associated with promoter, CGI and gene-rich regions (Figure 25). Those low efficiency origins do not associate clearly with a specific epigenetic mark except with the silencing histone modification H3K9me3 and with methylated DNA.

The second class associates at least two origins within a 14 to 4 kb window. This origin class, as well as the third class characterised by at least two SNS peaks located in a 4 kb window, associates more with early replicating regions, gene-rich sequences, CGI and Promoters. Class 2 origins were more frequently found localised in exons and introns with a quite relatively low origin efficiency. Class 2 correlated with active genes and enhancer marks, such has H3K4me1, 5hmC and H3K36me3. A strong correlation with the closed chromatin mark H3K27me3 was also observed for class 2 origins. Class 3 origins are strong origins found associated with CGI (45%) and with promoter (68%) (Figure 25). The class 3 associates with DNaseI hypersensitive sites (DHS) H3K9ac and H3K4me3, but also with transcription factors such as TATA-binding protein (TBP) or RNA polymerase II. Surprisingly, the polycomb complex, usually depositing repressive mark, has been found enriched at the class 3 origins level, giving bivalent histone marks at the origin level in pluripotent cells[70].

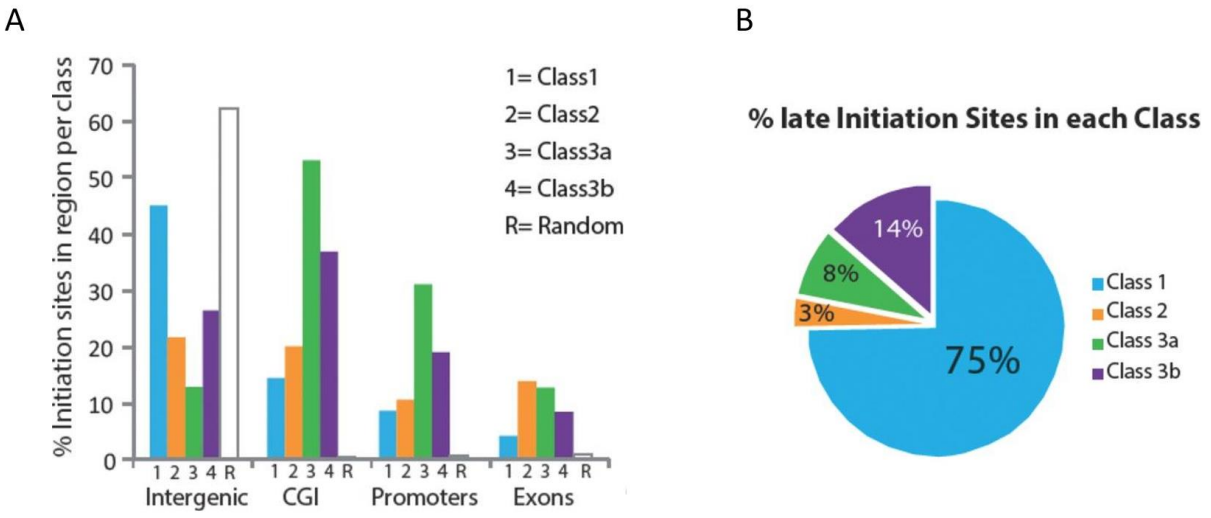


Figure 25

(A) Association of different classes of origins with genomic elements. (B) Distribution of origin classes in Late regions.

From Cayrou *et al*, (2015)

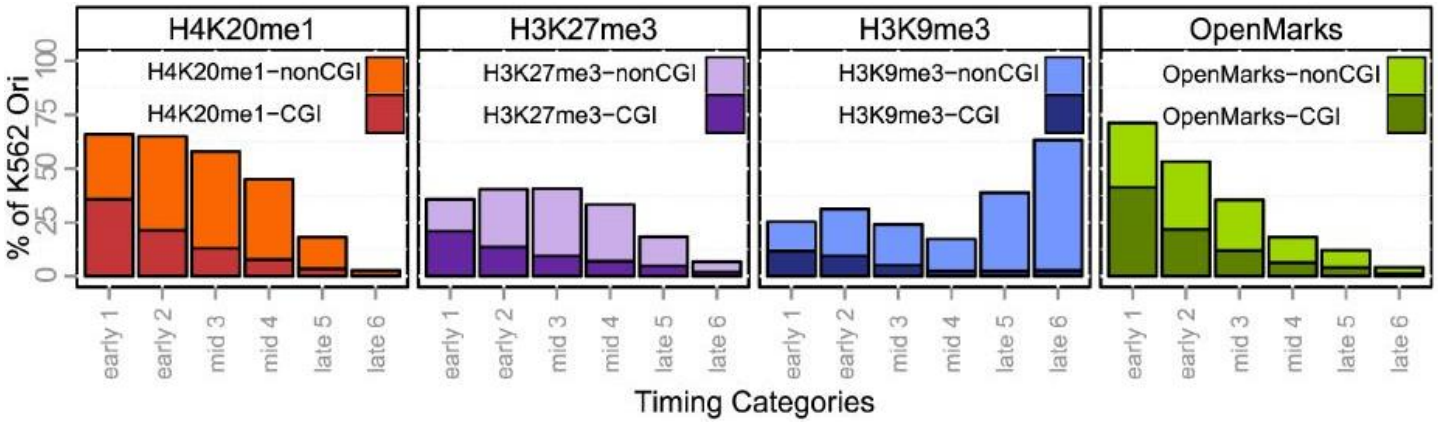


Figure 26

Histones marks enrichments at replication origins from early to late S-phase.

From Picard *et al*, (2014)

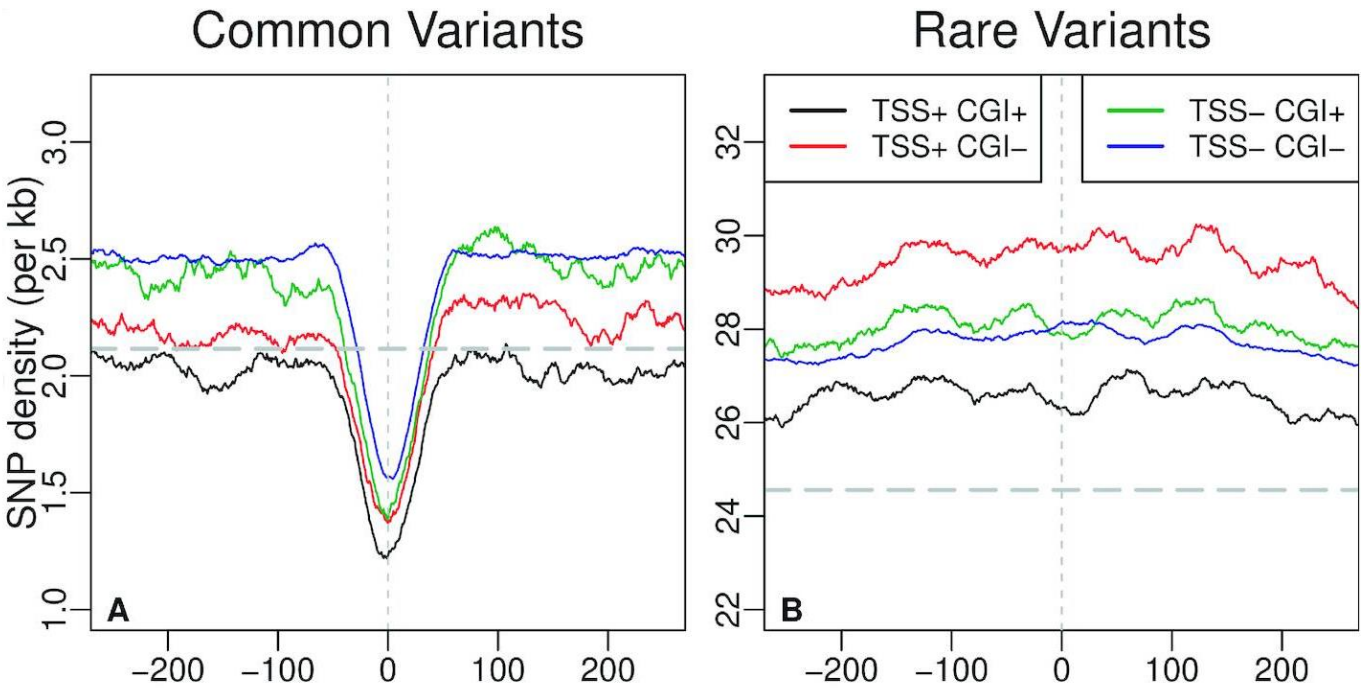


Figure 27

SNP density centered on SNS peaks in human using 1000 Genome project data.

From Massip *et al*, (2019)

iii. Replication origins and chromatin marks linked to replication timing program

To decipher the links between replication origins regulation and the RT program, replication origins have been sorted according to their replication firing timing, origins characteristics as well as the chromatin marks associated to them.

Early replicating origins are enriched with pG4 and GC-richness. Early origins are also more associated with CGI (32%) than mid-replicating origins (15%) and late origins (8%)[56]. Interestingly, early replicating domains contains more constitutive origins (67%) compared to late regions enriched with cell-specific origins. Histone post-translational modifications H3K9ac, H3K4me3 and the H2A.Z histone variant were found associated with early replicating origins. The H4K20me1 mark was found within fifty percent of origins that are mostly early and mid-S replicating origins. The chromatin mark, H3K27me3 deposited by the polycomb complex was also enriched, to a lesser extent (40%), with early and mid-S phase replicating origins. At last, late replicating origins were associated with the H3K9me3 histone mark, characteristic of heterochromatin, this mark was significantly depleted in early regions (Figure 26).

iv. Organisms comparative origins study

Replication origins in metazoan appear to be very dynamic and present some differences according to the organisms. A recent study on origins mapped with the SNS method in mouse, human and chicken reveals that replication origins were more dispersed in mouse compared to human and chicken. Indeed, 5% of the genome concentrates 17% of SNS peaks in mouse compared to 39% found in human and chicken cells[69]. pG4s were detected in the vicinity of SNS peaks. However, in human, pG4s were found mostly localised 150 bp away from the SNS peaks whereas, in mouse and in chicken pG4s were closer (50bp upstream).

Taking advantages of the high number of sequenced human genomes (1000 genome project) and the identification of rare and common nucleotides variants, it was possible to look for evolutionary constrains on genetic elements at the SNS peaks. A 40 bp region, centered on the SNS peaks appears to have a strong deficit in common variants (ancient mutations) that is not observed for rare variants (recent mutations)[69] (Figure 27). These results indicate that SNS peaks sequences undergo mutations at similar levels compared to the rest of the genome, but those mutations are not fixed during evolution, signing a purifying selection at the SNS peaks level. Moreover, this purifying selection is independent of replication origins associated functional elements such as CGI or promoter[69]. Such **selection pressure on a sequence indicates that essential cis-regulatory elements are found inside this 40 bp window.**

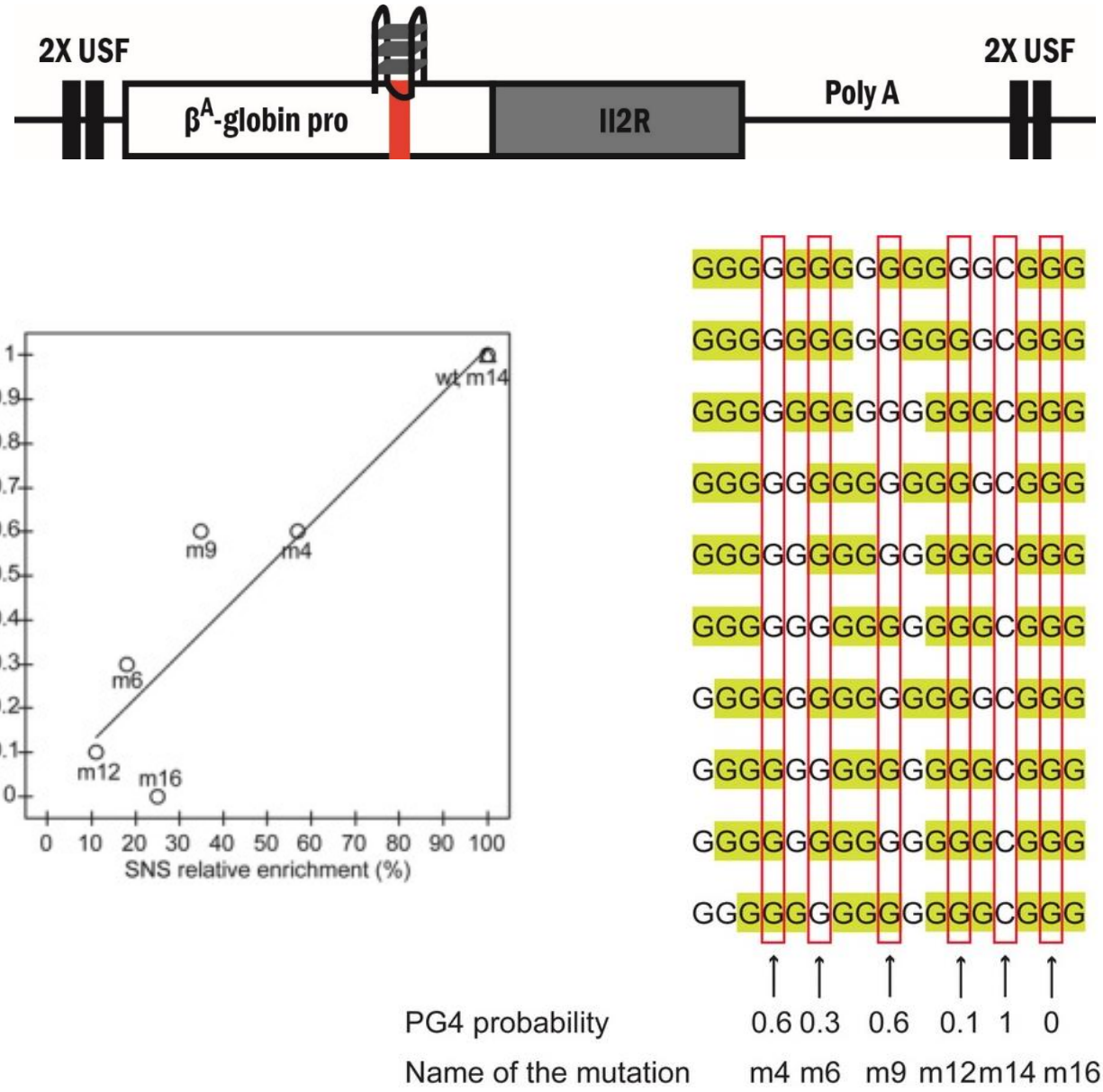


Figure 28

Investigation of the G4 formation role in replication origin activity. Red rectangles represent the mutated base (G to A) with the pG4 probability formation of the mutant compared to the WT sequence. Gs underlined in green represent the triplets of Gs involved in the 10 different potential G4 structures adopted by the WT sequence. On the right, the pG4 probability is plotted against origin activity defined by SNS enrichment for each mutant.

From Valton *et al*, (2014)

Search for **motifs enrichment led to the identification of several motifs common in human, mouse and chicken** such as CCC (GGG), GAG (CTC) and CG. C triplets could be involved in G4 formation, but according to the pG4 density (based on canonical G4 motif) near the SNS peaks it seems unlikely that these CCC motifs are involved in G4 formation[69]. However, using canonical G4 sequence strongly limit pG4 prediction (See section "evidence of G4 formation *in vivo"* for more detail). A previously identified motif in mouse was the AC (GT) motif[70] mostly enriched in late replicating origins, this motifs was also found enriched in human in this study.

The motifs' role has not yet been understood. One possibility would be on the recruitment of factors, or on the DNA double helix destabilisation as proposed in Drosophila[66]. Replication origins were enriched with specific DNA shape signatures such as reduced helix twist, propeller twist, minor groove and twist[66].

c. Genetic identification of *cis*-regulatory elements involved in replication origins activation

Before the advancement of genomic studies, genetic approaches to identify replication origins in metazoan have been used and led to the identification of several origins such as the human β^A-globin origin localised in the β-globin locus[51,71], firstly defined as an 8 kb sequence able to support DNA replication. Another well-studied model origin is the one found inside the chicken β^A-globin promoter. Retrospectively, a closer analysis of the mapped chicken β^A-globin origin validates previously identified characters associated with strong replication origins genome-wide. The chicken β^A-globin promoter sequence has been isolated and ectopically inserted inside the DT40 chicken cell line genome. This sequence of 1,1 kb long could sustain DNA replication at the ectopic position. It contains two pG4s, on opposite strands, deletion of one pG4 (GGGGGGGGGGGGGCGGG) disrupts the origin activity, such deletion could be mimicked by a point mutation that strongly alters G-quadruplex formation. Several pG4 point mutations altering the G-quadruplex formation to different extent were tested and the different degrees of alteration correlated well with the origin activity[72] (Figure 28). **Altogether, these results suggest a critical role of G-quadruplex formation rather than only the presence of a G-rich sequence.** Moreover, this study showed that pG4 inversion affected the localisation of the SNS peak from downstream the β^A-globin origin sequence to upstream, such result has also been observed in human[72,73]. Despite the central role of the pG4, deletion of a 250 bp region 3' of the pG4 drastically reduces the origin activity. Similar experiments have been made on a second replication origin containing two pG4s on the same strand at the endogenous locus. Again, pG4s were shown to be essential for the replication origin activity. Moreover, the two pG4s had a synergic activity, deletion of

both pG4s was necessary to disrupt the origin activity[72]. Investigation of the role of pG4 in origin function in mouse, human and Xenopus has been done more recently. In mouse, pG4 deletion on one model origin decreases the SNS enrichment. In human cells, transmission from mother cells to daughter cells of an episomal DNA containing the EBV origin partially deleted OriP was used. It showed that a 500 pb mouse Ori sequence containing an OGRE motif facilitated the replication of the episome[73]. The addition of a G4 stabilising drugs (PhenDC3) to mESCs in culture did not affect 77.9% of replication origins even though they were associated with pG4 motifs. The remaining 32.1% replication origins were either suppressed (5.1%), reduced (0.7%), reinforced (0.6%) or new (15.7%). Suppression and reduction of origin efficiency, even though the pG4 were stabilised, could be explained by the arising of new replication origins that "displaced" replication origins and made the suppressed and reduced origins unnecessary to fire[73].

Finally, the authors used the *Xenopus Laevis* High-speed egg extract (HSE) that is able to sustain replication origin licensing but not firing unless supplemented with low-speed egg extract (LSE). Nuclear DNA supplemented with HSE could efficiently form Pre-RCs on DNA and required the supplementation of LSE to have active DNA replication. The addition of competitor oligonucleotides containing pG4 sequences did not affect origin licencing but strongly diminished origin firing, suggesting that pG4s play an important role in origin firing factor recruitment in this model system.

These studies genetically confirmed the previously identified characteristic of replication origins genome-wide and the central role of pG4 on origin activity.

5. G-quadruplex

a. G4 structure

DNA *in vivo* has long been thought to remain organised into a canonical double helix form as described in 1953. However, some studies revealed the ability of some sequences to structure into un-conventional DNA structures. In 1962, structural study using X-ray diffraction showed that fiber formed by guanylic acids assembled in vertically stacked hydrogen-bonded guanine tetrads. Gellert *et al*. also reported the high stability of the structural arrangement provided by Hoogsten hydrogen bonds instead of Watson-crick hydrogen bonds[74]. **Those four guanines organise themselves in a plan called G-tetrad (or G-quartet), stabilised by monovalent cations such as sodium and potassium (Figure 29A). G-tetrads stacking on top of each other lead to the formation of a G-quadruplex**.

G4 structures are very dynamic, and several conformations can be adopted by most G4 motifs. G-quadruplex can be intramolecular (formed from a single DNA molecule) or intermolecular (formed

from at least two DNA molecules), the already described G4 canonical motif is mostly true for intramolecular G4. Indeed, intermolecular G4 does not necessarily require four tracts of Gs, it can be only two tracts on one strand and two others on the other strand. Moreover, strands involved in the G4 can be parallel strands, anti-parallel strands (2 parallel strands and 2 anti-parallel strands), or hybrids (3 parallels strand and 1 anti-parallel strand)[75,76] (Figure 29B). Another layer of complexity for the G4 relies in the loops composition and size, indeed loops have been defined as being one to seven base pairs long, but longer loop can also be observed in G4 structure. Loop size and composition have been shown to affect the G4 conformation. Single-nucleotide long loops tend to form parallel G4 and sustain a parallel G4 organisation with one long loop and two single-nucleotide loops[77]. *In vitro*, loop size does not affect G4 formation, but with short loop, G4 rather structures themselves in parallel form whereas longer loop tends to structure in anti-parallel of hybrids form when only thymidine bases are found into the loops. Moreover, when loop size increases the G-quadruplex thermal stability decreases[78].

Independently of loop flexibility, tract of Gs can be disrupted by other bases and can still form a G4. However, resulting G4s exhibit a bulge that connects Gs from two other G-tetrad instead of connecting Gs from other columns as for loops (Figure 29C). However, bulges composition affects the G4 stability, Cytosine and Thymine giving similar G4 stability and Adenosine giving a more drastic G4 destabilisation. Bulge size and position, as for base composition, does affect G4 stability but the G-quadruplex are still able to structure themselves *in vitro*[79].

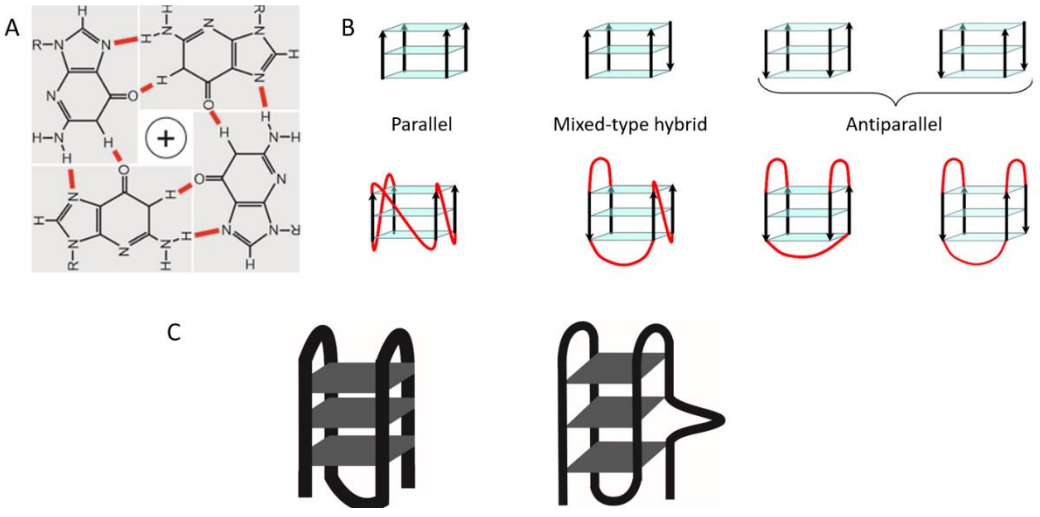


Figure 29

A. Organization of G-tetrad (G-quartet) with in the middle a cation, red segments are Hoogsten bounds. B. Overview of possible G4 structure organizations. Arrows indicate strand direction from 5' to 3', parallel, hybrid and antiparallel structures are shown. Red loops are possible organization to conserve G4 structures. C. cartoon of possible canonical G4 (left) and G4 with a bulge (Right)

From Bugaut *et al*, (2008) and Rhodes *et al*, (2015)

b. Evidence of G4 formation *in vivo*

G4 structure has been easily observed *in vitro*, but G4 existence *in vivo* has long been questioned. To answer this simple question, several approaches have been used. A first indirect approach relies on the mutation of helicases known to resolve G4 structure *in vitro*, those helicases called WRN, BLM and FANCJ are important helicases which mutations induce genome instability. WRN helicase has been found associated with telomeres that are structured with G4s to protect chromosome ends from degradation. WRN mutation causes Werner syndrome, characterised by a telomere shortening. For BLM, FANCJ and other helicases such as Pif1, their mutations (causing syndromes as the Bloom syndrome and Fanconi anaemia) induce genomic instability and the loss of large DNA fragments localised at predicted G4 formation sites[76].

More direct evidences arose from the use of antibodies directed against G4. Several antibodies were developed to directly address *in vivo* G4 formation. A first study generated a single-chain antibody Hf2 which recognises an intramolecular G4 presenting a canonical sequence able to fold in a parallel and anti-parallel way[80]. This antibody has been used to do ChIP-seq analysis and allowed the identification of 768 peaks[81]. However, ChIP-seq analysis only identified a subset of G4 (370 000 putative G4 are found in the human genome[82]). Among the identified G4 by ChIP-seq, only 175 peaks contained the canonical G4 motif, underlining a great G4 plasticity *in vivo*. These identified peaks only represent a specific population of G4 since the antibody was generated against a specific structure and could have been subjected to artificial G4 formation during chromatin purification. To increase the G4 detection sensitivity and to control for the proper *in vivo* G4 formation different antibodies and different experiments have been done. The BG4 antibody used in Fluorescent in Situ Hybridisation (FISH) led to the direct observation of discrete G4 structures at the genome level[83]. The BG4 antibody could detect a large panel of G4 structures, going from intramolecular parallel/anti-parallel/hybrids propeller G4s to intermolecular G4s. The **polyvalent BG4 antibody allowed visualisation of G4s at telomeres as expected, but most observed G4s (75%) were found located inside the chromosomes**. The number of G-quadruplex detected increased during S-phase and with the use of G4 stabilising drugs[83].

Regarding the dynamic folding of G4s and the high number of potential structures it appears clearly that the G4 canonical motif used in bioinformatics research only identify a small subset of pG4. To outcome this limitation, a very elegant approach has been used that took advantage of the illumina sequencing technic and the ability of structured G4 to induce DNA polymerase blockage. Human genomic DNA has been sequenced in disfavoured G4 formation condition and promoting G4 formation condition (K+ or Pyridostatin (PDS) a G4 stabilising drug). Comparison of the base calling quality in G4 structure condition (low base calling quality after G4 encountering) with un-structured G4 condition (high quality base calling) allowed the identification of formed G4[84]. Around 700 000 G4 were identified

by the G4-seq technic which was two times more than what was predicted. Interestingly only 200 000 (in K+ and PDS conditions) observed G4 were computationally predicted and present a canonical G4 motif. However, the majority of G4 presents long loop (>7 nt) or a bulge, with an average number of identified G4 of 143 000 and 163 000 respectively. A third G4 class defined as others with 100 000 observed G4, containing two G-tetrads G4 or even long loops with a simultaneous bulge G4 (Figure 30). The G4-seq, a high-throughput *in vitro* analysis of G4, allowed the refinement of the previously used G4 computational research algorithm but mostly, underlined the **presence in the genome of various sequences able to structure themselves into G4** (at least *in vitro* in physiological conditions) **that could not have been observed with the previous technics**.

Recently, a second ChIP-seq study made with BG4 antibody identified around 10 000 G4 peaks in human cells (mostly canonical G4)[85]. G4 were found enriched 1kb upstream of TSS and at the 5'UTR of genes which was also expected by computational analysis[86]. However, computational analysis found a G4 enrichment at gene first intron localised at the exon-intron junction that was not found back in the ChIP-Seq analysis[85,87].

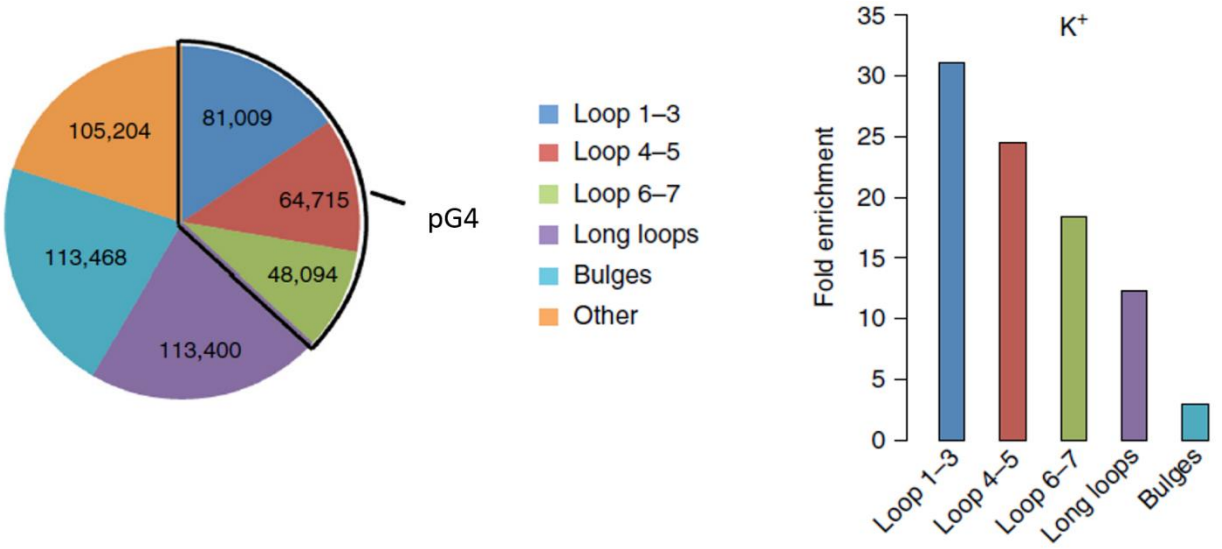


Figure 30

Results from G4-seq indicating G4 distribution according to their loop sizes or presence of a bulge. pG4 (canonical predicted quadruplex).

From Chambers *et al*, (2015)

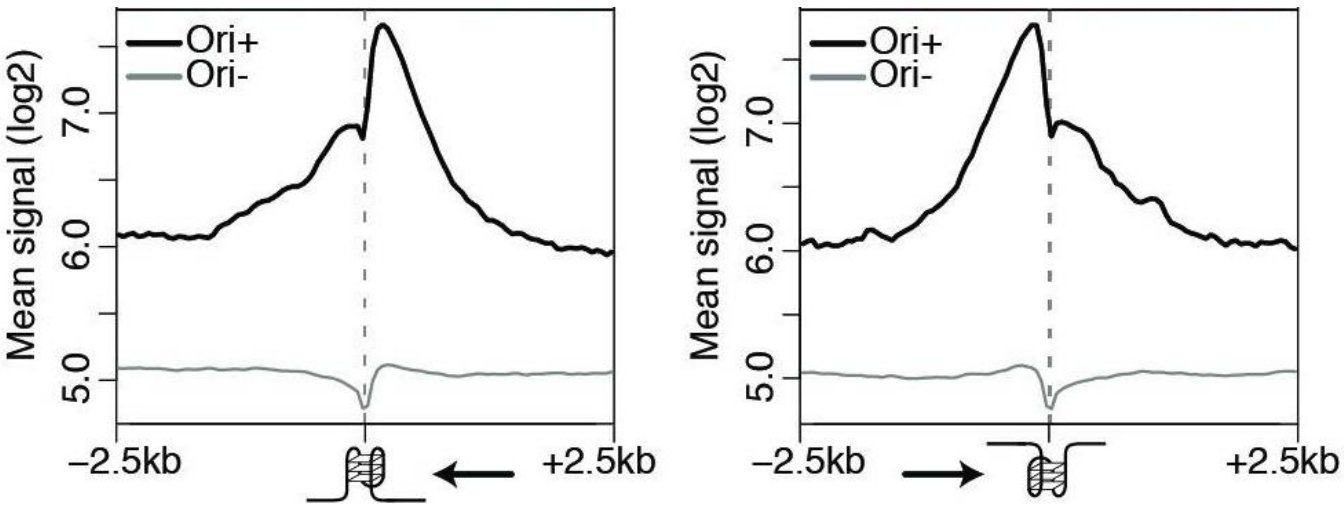


Figure 31

Profiles of SNS enrichments at G4 associated origins with respect to G4 orientation. Black arrows indicate the replication fork direction.

From Comoglio *et al*, (2015)

c. G4 structure resolution and conflicts at the replication fork

The mapping of SNS of different sizes in Drosophila cells suggested that **structured G4 have the ability to transiently block the replication fork.** This observation made at high throughput level[66] was firstly made by qPCR analysis of model origins in chicken cells[72]. Indeed, the SNS enrichment drops drastically at the position of the G4 whereas the SNS signal on the opposite fork does not have this shape (Figure 31). Cells however can naturally resolve these G4 structures with the help of helicases.

In S. *cerevisiae*, human minisatellites CEB1 insertion was used to test their impact on genomic instability. This minisatellite is a 39 bp motif repetition able to form a G4 structure containing two one-nucleotide loops and one four-nucleotides loop. A 1,8kb long CEB1 minisatellite repeats induces genomic instability in a G4 stabilisation context or in a Pif1 deleted mutant. Pif1 has been involved in G4 structure resolution *in vivo*[88]. Further study using this minisatellite showed that fork progression and genomic stability are impaired when the G4 structure is located on the leading strand template but not on the lagging strand template[89]. Moreover, the replacement of the CEB1 minisatellite by CEB25 abolishes genomic instability in both conditions (G4 stabilisation or Pif1 deletion). The CEB25 G4 present two one-nucleotide loops and one long loop of nine nucleotides. CEB25 minisatellites insertion only exhibits genomic instability if the long loop is reduced to at least four nucleotides, which result in an G4 thermal stability increase[90].

Interestingly, replacing CEB25 G4 by other described G-quadruplex does not necessarily induce a genome instability and this, even though the G4 is stable *in vitro* (e.g. c-myc G4). Moreover, helicases do not seem to have similar targets, as the deletion of different helicases does not impaired the genomic stability[90].

In addition to genomic stability, replication **fork pausing due to G4 can induce uncoupling between DNA replication and recycling of old histones**. Indeed, when the DNA replication machinery progresses, parental histones are removed from the DNA and then evenly re-distributed between the sisters chromatids with new histones. Replication fork and histone recycling decoupling lead to the progressive loss of chromatin marks inducing an epigenetic instability that will lead to the de-repression of gene or loss of transcription for transcribed genes.

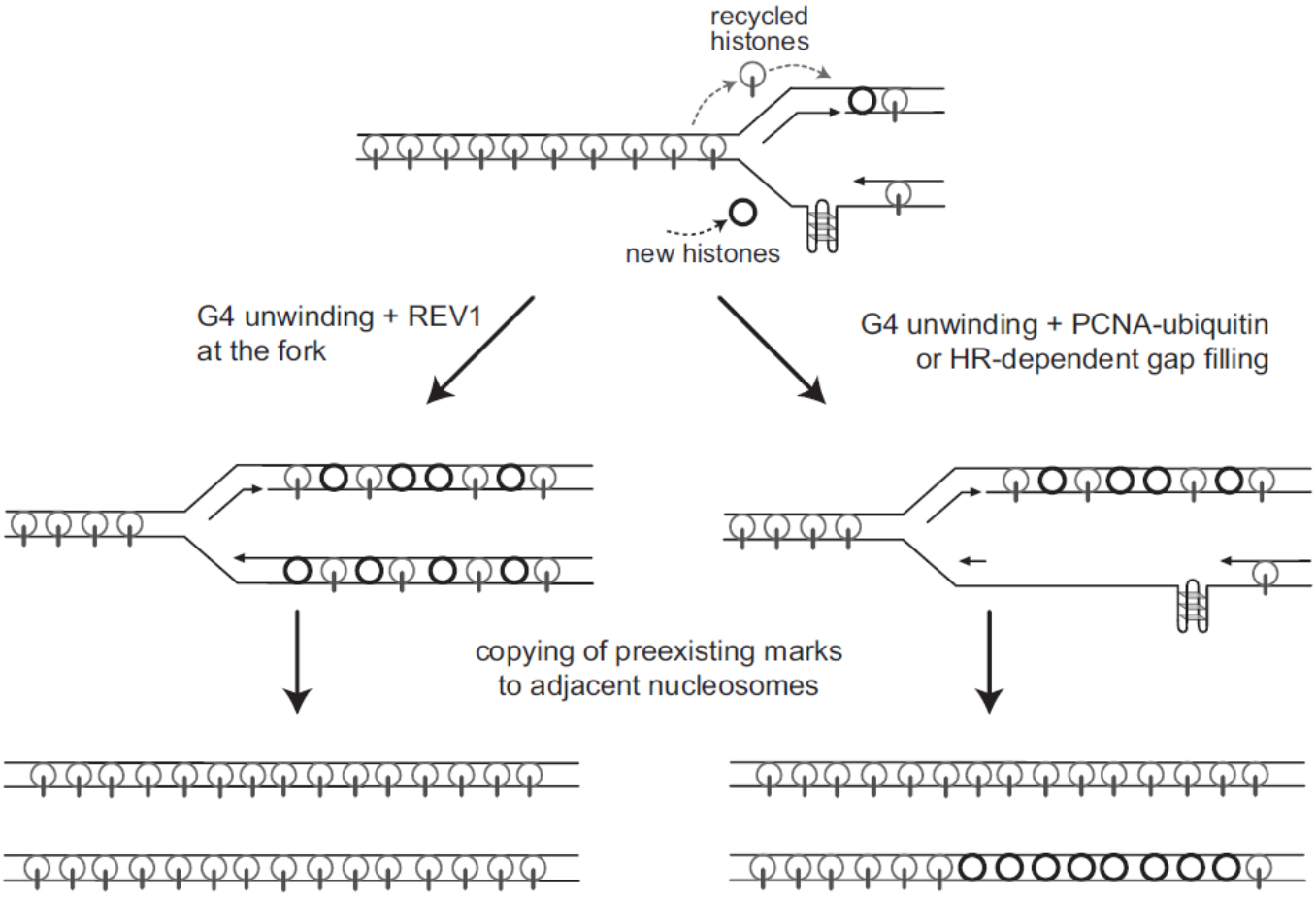


Figure 32

Illustration of parental histone marks recycling in a WT condition (left) and when histone recycling and replication fork progression are uncoupled in mutant condition (right).

From Sarkies *et al*, (2010)

In DT40, an elegant study focused on the Y family DNA polymerase Rev1, able to maintain fork progressing on damaged DNA through the activity of several polymerases, its deletion leads to resolving of DNA damaged or gaps in a post-replicative way. Rev1-deficient cells can replicate damaged DNA but this replication is dissociated from the moving forks and so of old histones recycling. In DT40 cell lines Rev1 deletion induces transcription activation and repression of several genes[91,92]. To address the role of G4 on the uncoupling of DNA synthesis and histones recycling, the authors first studied the ρ-globin gene found inside the β-globin locus and containing three pG4 in the second intron. Histone mark have been extensively studied previously and in the DT40 cell line lymphoid background, the repressed ρ-globin gene exhibits H3K9me2 chromatin mark. In a Rev1 mutant context, H3K9me2 was lost and led to a 100-fold gene expression increase of the ρ-globin gene [91]. A replicating plasmid assay in a Rev1 mutant context revealed a DNA replication defect only when the G4 was located on the leading strand template. The role of the ρ-globin G4 has been tested by adding it 500bp downstream of the TSS of the Lysozyme C gene (LYSC). Its expression did not change upon Rev1 deletion and its gene core did not present any detectable G4. This gene is repressed in the DT40 cell line and exhibits H3K9me2 histone marks. However, ρ-globin G4 in Rev1 mutants did affect LYSC gene histone mark and its expression level is increased by 30-fold[91]. Genomic analysis of transcription-affected genes upon Rev1 deletion revealed that 71% of upregulated genes have G4 motif within a 3kb window around TSS. Deletion of Rev1 does not only lead to upregulation of genes but also to downregulation of genes such as CD72 that present a G4 containing five tracts of 3-4 Gs and loops of 3-7 nucleotides located in the third intron. CD72 gene showed loss of H3K4me3 and H3K9/14ac in Rev1 deleted DT40 mutant. Similar results could be observed with the BU-1 gene, containing a G4 at the end of the second intron located 3,5kb downstream of the TSS. Rev1 deletion led to the loss of H3K4me3 and H3K9ac and H4ac[92]. Interestingly Bu-1 genes expression can be easily followed thanks to Bu-1 proteins localisation at the cell membrane (Figure 32).

Deletions of G4 resolving helicases WRN, BLM and FANCJ induce a decrease of gene transcription. Interestingly, the subset of genes affected by BLM and WRN deletions is quite low (2284 and 304 respectively) and BU-1A gene does not show any loss of expression. However, when both BLM and WRN are mutated, BU-1A expression is drastically lost and the total number of affected genes reaches 6152[92], dysregulated genes are associated with G4 motifs. Such combinatory mutations effect underlines a redundancy in the targets of BLM and WRN helicases. FancJ deletion alone leads to the dysregulation of 7152 genes highly associated with G4 motifs including the BU-1A gene. However, BU-1A repression was not linked to genetic alteration.

To investigate the role of G4 itself, the Bu-1A gene has been used as a reporter for epigenetic instability. In a Rev1 deletion mutant context (as well as FANCJ, WRN/BLM deletions mutants) the Bu-1A gene undergoes epigenetic instability leading to its downregulation. Bu-1A gene is located inside an early replicated domain, with no replication origin within the gene; it is passively replicated by origins located at locus both ends. In order to confirm the role of G4 formation on the epigenetic instability, two point mutations impairing the G4 formation *in vitro* have been made inside the endogenous G4 found 3.5 kb downstream of the TSS. Those mutations led only to a small expression level loss compared to WT G4 confirming the role of the G4 on the gene downregulation[93]. G4 motif inversion did not result in a BU-1A expression loss, confirming that G4 impairs fork progression when located on the leading strand. Moreover, as discussed before, G4s have a very dynamic range of folding associated with variable thermal stabilities. To assess the impact of G4 stability, different G4s have been inserted 3,5kb downstream of the BU-1A promoter. Four G4s have been studied: one parallel folded G4 of two stacks, one antiparallel, two hybrids G4 and the ρ-globin G4 containing five G stack but with a median melting temperature (63°C). Interestingly, the G4 with the highest melting temperature (>95°C) did not have a strong effect on the epigenetic instability whereas the G4 with the lowest melting temperature (27.6°C) had a strong one. Nonetheless, a G4 with a similar Tm (temperature melting) did not have the same effect on the BU-1A expression level. In conclusion, there is no clear evidence of a direct correlation between the degree of gene expression loss and the Tm of the G4 measured *in vitro*. It rather seems that the sequence of the G4 loops is more relevant in predicting its action on epigenetic instability. Finally, moving the G4 motif downstream of its initial position decreased its action on the gene transcription. Beyond a critical distance of 4,5kb from the TSS BU-1A expression was no longer affected in a Rev1 deleted context[93].

All these studies suggest a deleterious role of G4 inside genomes and especially at the proximity of gene promoters. It remains thus a mystery why about 40% of human promoters are associated with pG4[86].

d. G4 potentially bound proteins.

After the confirmation that G4s are actually formed *in vivo*, interest has shifted toward the identification of proteins able to recognise these structures. Such proteins carry a RGG/RG domain such as the nucleolin or Ewing's sarcoma protein (EWS). The RGG domain of EWS can bind RNA and DNA G4s with a dependence on the loops size rather than on its composition. Interestingly, EWS binding turns antiparallel G4 into parallel one[94]. Nucleolin is a nucleolar phosphoprotein and plays an important role among others in ribosome biogenesis. Nucleolin can bind the C-myc G4 motif as other proteins (such as TTF-I, TFII-I…) involved in part in chromatin remodelling, transcription, RNA splicing etc. Nucleolin binding affinity was greater for the c-myc G4 than for its RNA substrate[95]. Moreover,

nucleolin can recognise the c-myc G4 unfolded and induce its folding *in vitro*[95]. Interestingly, nucleolin exhibits a preference for G4 with at least one long loop (more than 3nt) compared to one-nucleotide loops or three long loops[96].

A protein more directly related to replication, Rif1, was proposed to recognise G4. ChIP-seq against Rif1 in the yeast S. *pombe* localised enriched peaks near late and dormant-origins. Next to Rif1 identified peaks, G-rich motifs were found as repeats (mostly two) capable to structure themselves into G4 *in vitro*[97]. This G-rich motif only contains a limited number of Gs that does not allow it to fold into a G4 structure. Interestingly, at most Rif1 binding sites, two G-rich motifs were found frequently oriented in a head-to-tail fashion. Mutations of those Rif1 sites led to the Rif1 interaction loss *in vivo.* Rif1 consensus sequences *in vitro* characterisation revealed the ability of these sequences to structure into G4. Moreover, S. *pombe* Rif1 purified protein bound G4 *in vitro* with higher affinity than other DNA structures such as cruciform DNA and replication fork[97].

Among the protein involved in origin firing, MTBP and RECQL4 were both found to interact with G4[98,99]. MTBP interacts with Treslin allowing the recruitment of CDC45, a protein essential for Pre-IC formation (Figure 8). MTBP Interaction with Treslin forms a hetero tetramer complex constituted of two Treslin and two MTBP, with no free-Treslin or free-MTBP found. Dissection of MTBP revealed the essential role of its C-terminal domain in DNA replication. This domain can bind double stranded DNA (random 30-mers). The binding to dsDNA can be efficiently competed with single stranded 30-mers capable of forming G4 but not with random ssDNA 30-mers. Its deletion led to a defect in DNA replication mediated through the lack of MTBP interaction with DNA and the impairing CDC45 recruitment to Pre-IC. Indeed, MTBP knockdown cells exhibit a longer S-phase linked to a decrease in origin firing detected by an increase in inter-origin spacing[98]. *In vitro* investigation of RECQL4 N-terminal domain identified three DNA binding domain (Homeodomain-like, a SSDNA binding domain and recognising Y-DNA, ssDNA and dsDNA) and exhibited a strong affinity for structured G4[99].

Finally, ORC is the first complex loaded on replication origins. By contrast with yeast ORC, metazoan ORC does not exhibit any sequence specificity. However, ORC has a sequence independent ability to bind to dsDNA (40-mers) and single stranded DNA (ssDNA), wheras ssDNA $(GGGTT)_8$ and $(GGGAA)_8$ compete efficiently with dsDNA, the ssDNA $(AACCC)_8$ cannot. Moreover, ssDNA $(GTTGT)_8$, $(GTTTT)_8$, $(GTGTGT)_7$ do not compete with ORC-dsDNA complex. SsDNA $(GGGTT)_8$ and $(GGGAA)_8$ have the propensity to get structured into a G4. Finally, ORC tend to have a slightly better affinity for ssRNA than ssDNA when the same motifs are tested[100].

The affinity of ORC for ssDNA depends on the ORC1 sub-unit. Its C-terminal part (413-511) confers the ssDNA binding specificity to ORC. The difference of affinity for ssDNA and ssRNA observed with ORC is lost when only the ORC1 C-terminal part is tested [100].

From these studies, it is now accepted that G4s are very dynamic structures that can form *in vivo*. G4 are involved in several molecular processes such as telomere protection, transcription and translation regulation. Its major role in DNA replication has emerged during this last decade. The discovery of proteins involved in DNA replication initiation that can recognise **G4s, together with the finding that G4s are key *cis*-elements at replication origins, should allow in the near future a clear understanding of how replication is initiated in complex vertebrate genomes.**

6. Nucleosome positioning on replication origins

As mentioned before, DNA is associated with histones to form nucleosomes. The sequence size of the DNA wrapped around the octamer of histones is 147bp long, DNA "enters" and "exits" the nucleosome core particle at the same point. The middle of the wrapped sequence is called the dyad. Nucleosomes are very dynamic structures and can be spread over the genome in an un-specific manner. However, nucleosomes can also be well positioned on DNA in several conditions.

At a specific region nucleosome organisation can be defined by two criteria, nucleosome positioning and nucleosome occupancy. Nucleosome positioning can be defined as the probability in a given cell population of having the dyad localised at a specific genomic position. A region with a high positioning is characterised by a nucleosome having its dyad located at the same genomic position in most of the cells whereas in a region of low positioning nucleosome dyads are spread along the sequence[101]. Low positioned nucleosomes can also be named fuzzy nucleosomes.

Nucleosome occupancy corresponds to the number of nucleosomes localised inside a genomic window in a cell population. A poor nucleosome occupancy is defined as the presence of nucleosomes only in a subset of cells whereas a high nucleosome occupancy is characterised by the presence of nucleosomes in a certain window in most cells[101].

Nucleosomes can be well positioned with high occupancy or be not well positioned with low occupancy, but they can also be well positioned with a low occupancy or be not well positioned with a high occupancy. Moreover, another feature observed at the genomic scale is that **some regions are constitutively free of nucleosome (NFR),** most of them being at promoters. However, some regions have labile nucleosomes that can be removed during gene activation (stress-regulated genes). These regions are named nucleosome-depleted regions (NDR) even though the term NDR is often used interchangeably with NFR.

How nucleosomes are well positioned *in vivo* is not clear. It is now well established that nucleosome organisation involves contributions of several factors including the DNA sequence and the action of transcription factors together with chromatin-remodellers. *In vitro*, nucleosomes tend to position themselves depending on DNA sequences[102]. High throughput data in yeast S. *cerevisiae* revealed that nucleosome positioning does not rely on a particular base sequence but on the bases properties[103]. Indeed, nucleosomes tend to be positioned on sequences exhibiting "weak" bases (such as As and Ts) and "strong" bases (Gs and Cs) occurring in a 10 bp periodicity with weak and strong bases shifted by 5bp. This sequence is predicted to allow the DNA major groove expansion and to contract the minor groove. It is believed that the DNA bendability is involved in nucleosome positioning[104].

However, chromatin remodellers affect *in vivo* nucleosome positioning, independently of the sequence. Several families of chromatin remodellers exist, the SWI/SNF family, the ISWI and the SWR1. The SWI/SNF chromatin remodellers can both slide and eject nucleosomes from DNA and their functions are often associated with transcription activation. The SWI/SNF chromatin remodellers have domains that bind acetylated histone tails promoting their targeting towards promoters undergoing activation. ISWI family in opposition to SWI/SNF family tends to slide nucleosomes with a fixed linker size and to target non-acetylated histone tails. The ISWI chromatin remodellers family is involved in gene silencing[105]. The SWR1 family is involved in the substitution of canonical H2A histone protein by H2A.Z histone variant at the NDR of active promoters. Histone H2A replacement by H2A.Z alters the histone stability at the promoter and could play several roles on the transcription regulation[106].

Histone positioning actors *in vivo* have been identified, but their clear interplay remains elusive. Interestingly, at promoter, histone position can affect the gene expression. Two kind of promoters can be defined: open promoters and covered promoters, corresponding to constitutive and highly regulated genes respectively. Open promoters, as defined by their names, have a NDR containing transcription factors binding sites. These accessible sites allow transcription factors binding and recruitment of RNA polymerase II. It has been suggested that such molecular crowding and protein recruitment could affect the nucleosome positioning[104,105]. By contrast, covered promoters have nucleosomes deposited over the promoter, however, when needed, the gene can be transcribed after the opening of the promoter. However, most of the transcription factors binding sites are bound by nucleosomes. Pioneer transcription factors can yet recognise unbound binding sites found in linker DNA between nucleosomes but also nucleosome-bound recognition sites, if binding required only one face of DNA such as for glucocorticoid receptor. Pioneer transcription factor recruitment induces the promoter opening and NDR formation. An important feature of covered promoters is the presence of

a TATA-box at the promoter, whereas most open-promoters do not have one. However, all RNA-polymerase II dependent transcription requires the central transcription factor TATA binding protein (TBP). At covered promoter, TATA-boxes are located at the edge of nucleosome providing partial blockage that requires chromatin remodelling to ensure transcription activation[105].

Active open promoters exhibit NDRs of variable sizes and flanked by well positioned nucleosomes with high occupancy (depending on gene expression). These well-positioned nucleosomes can also be defined as phased nucleosome arrays. Nucleosomes downstream NDRs are defined as +1, +2, +3 etc… the +1 being the first nucleosome flanking the NDR. Nucleosomes upstream the NDR follow the same logic and are named -1, -2, -3… NDR flanking nucleosomes (+1 and -1) are the most strongly positioned nucleosomes and nucleosome positioning, slowly decays. Studies suggest that sequences, molecular crowding and chromatin remodellers define the first positioned nucleosomes and that following nucleosomes are positioned statistically due to molecular constraints, such theory could explain the slow decreasing of nucleosome occupancy when moving away from the NDR[104,105]. NDRs and +/-1 nucleosomes after DNA transcription are rapidly phased at their initial position, however, +2, +3 etc nucleosomes are positioned back more slowly[107].

a. Technics to study Nucleosome positioning.

To study nucleosome dynamics, several genomic assays have been developed. One of the most common, the ChIP-seq, uses antibody directed against histones or their modifications to immunoprecipitate the associated DNA which is then sequenced. This technic relies on the physical fragmentation of DNA by sonication after the use of formaldehyde to trap histone-DNA interactions *in vivo*. This method unfortunately only gives a very partial and un-precise map of nucleosomes, but provides a dynamic view of nucleosomes that can be immunoprecipitated according to the histone variants and post-translational modifications.

Another common technic used to map the nucleosome relies on the enzymatic digestion of un-protected DNA, the linker DNA located between nucleosomes. Histone associated DNA is protected from the enzymatic activity when mild digestion is applied. The enzyme used, the micrococcal nuclease (MNase), is an exo-endonuclease and has a bias toward weak bases (A and T) that are more easily digested than strong bases (G and C)[108]. Due to MNase bias, it has been hypothesised that NFRs are not devoid of nucleosomes but rather have weakly positioned nucleosomes that are more rapidly digested compared to neighbour nucleosomes, giving an artefact NFR. These MNase-sensitive nucleosomes have been proposed to localise at promoter level rather than true NFRs[109]. After more precise investigations, these MNase-sensitive nucleosomes appear to be non-histone

proteins[110].Despite MNase bias, use of a caspase-activated DNA having similar action on DNA as the MNase gives very similar results between the two datasets thus cross validating the two technics[111].

Another limitation of MNase-seq that must be taken into account is the lack of direct evidence of the protein nature involved in DNA protection. Indeed, after MNase digestion 150 bp long recovered fragments are mostly considered to be nucleosomes, but the DNA protection could be due to non-histone factors and/or complexes. MNase-seq is a quite simple technic to set on and allows to reach a more precise nucleosome positioning than with ChIP-seq. However, MNase-seq does not directly address the DNA protecting protein nature. This may be remedied by coupling MNase with a histone ChIP step. Chromatin is firstly digested with MNase reaching the mono-nucleosome scale, and then a histone ChIP step allows to clearly identifying the DNA protecting protein histone nature. However, the digestion deepness cannot be resolved with ChIP-seq coupling.

From 5 738 identified MNase-sensitive sites in yeast, only 2 631 were nucleosomes, the remaining sites were non-histone proteins. A focus on DNA sequence unravel the sequences A/T richness associated with MNase sensitive nucleosome that does not allow to clearly dissociate unstable nucleosomes or A/T rich MNase bias[110]. Time course digestions of chromatin by MNase in Drosophila revealed that indeed, nucleosome positioned at A/T rich sequences were digested by the MNase and that G/C rich sequences were digested only after longer incubations[108]. Most published MNase-seq data use deep digestions after which 80% of the material is cut into mono-nucleosomes. Deep sequencing of such samples then inferred the nucleosome positioning from the mapped peaks. However, **it has to be kept in mind that with such deep digestion, the absence of peak does not necessarily indicate the presence of a NFR** but could results from the preferential MNase digestion of an A/T rich sequence. Therefore, **MNase, in most of the cases, identifies bona fide nucleosome position** but has to be properly set and analysed in order to avoid false conclusions.

Another set of technics relies on a chemical reaction and requires the substitution of histone 4 Serine 47 into a cysteine. At this position, the cysteine is close to the nucleosome dyad DNA. The addition of sulfhydryl-reactive, copper-chelating label, N-(1,10-Phenanthrolin-5-yl) iodoacetamide catalyses the formation of short-lived hydroxyl radicals that results in DNA cleavage adjacent to the mutated H4 cysteine[112]. However, histone 4 modification leads to cleavage around the dyad, with a maximum length of 12bp, which is too short for a direct sequencing. Moreover, the chemical reactive can cleave un-protected DNA giving background. A similar approach relies on the substitution of histone H3 glutamine 85 into a cysteine. With this substitution the induced cleavage releases 51bp long fragments centred on the nucleosome dyads[113] and thus can directly be sequenced to address nucleosome positioning. The major limitation of these technics is the high number of histone H4 (or H3) coding

genes in metazoan. This method has been used successfully in S.*cerevisiae* where only one copy of histone coding gene is present. Another limitation of this approach is that histone variant, if not modified, will not lead to DNA cleavage and therefore will not be included in the study. It is important to note that such technics brought an unprecedented level of resolution for maps of nucleosome positioning with minimal bias. They confirmed the preferential positioning of nucleosomes at the weak/strong 10 bp alternating bases. Moreover, nucleosomes among a cell population tend to be well-positioned at precise points mostly at regions flanking NDR. Nucleosome dyads occupy a major position, with alternative minor position with a 10 bp pattern. This disposition may be linked to the helical twist. Mutation of the chromatin-remodeller RSC leads to loss of the major nucleosome positioning that then tend to dispatch over the different alternative positions. Finally, linker DNA has a dynamic range of length with a path of 10 bp (for example starting with 4bp and then increasing to 14bp, 24bp, 34bp…)[113].

Each method described above presents pros and cons but allows in any case to study nucleosome dynamics genome-wide. A great emphasis has been made on studying nucleosome dynamics at promoters but more recently the mapping of replication origins allowed to focus also on nucleosome organisation at these important sites.

b. Nucleosomes organisation at *S. cerevisiae* ARS.

To understand mechanisms involved in replication origin activity, the focus has been made on nucleosome organisation around replication origins. An *in vitro* study using the well characterised ARS1 origin showed that the strong positioning of a nucleosome over the ARS1 sequence led to the loss of origin activity[114].

ARS1 is constituted of the ACS sequence contained inside the A element and has a B1, B2 and the non-essential B3 sequence recognised by the transcription factor Abf1. *In vitro* nucleosome deposition on a plasmid containing the ARS1 led to a random nucleosome loading on the plasmid independently of the ARS1 sequence. The addition of ORC and Abf1 proteins *in vitro* induced their DNA recognition and fixation independently of each other. However when both proteins were added, a NFR was created over the ARS1 origin, such NFR was also found at the ARS1 sequence *in vivo*[115]. To investigate the role of ORC on nucleosome positioning, the A sequence has been mutated, resulting in -1 nucleosome invasion over the ARS1. Experiments with mutation of the Abf1 binding site led to similar results with the +1 nucleosome. However, the B2 sequence mutation did not affect nucleosome positioning. Pre-RC recruitment at the ARS1 origin has been investigated to check its role on the nucleosome positioning. As expected Pre-RC recruitment at ARS1 was impaired when A and B2 sequences were deleted resulting in no replication activity. However both mutant origins exhibit different nucleosome

behaviour that exclude the pre-RC as the major responsible actor involved in nucleosome positioning[115]. The ARS307 contains the A, B1 and B2 sequences and like many ARSs, lack the Abf1 binding site. However, ARS307 exhibits a NFR overlapping the ARS, that was dependent on ORC binding only[115].

Similar approaches at a genomic scale unravelled the differences in nucleosome organisation at active and non-active ARSs. Prior ORC ChIP-seq identified around 250 active origins and confirmed the presence of an ACS. The S. *cerevisiae* genome contains, by various metrics, 6000 - 40 000 potential ACS raising the question of why these elements are not functional origins. To explore the role of the ACS and potential surrounding in encoding an NFR, the chromatin landscape of 238 highest-scoring ACS motifs that were not within genes was analysed together with functional origins. MNase-seq analysis was performed over active (ORC-ACS) and non-active ACS or non-replicative ACS (nr-ACS). Active ARS showed approximatively 125 bp width NFRs with an asymmetrical distribution over the ACS (Figure 33). Indeed, ACS was found closed to the -1 nucleosome, DNase I footprint on the ARS indicated that 90 bp of the NFR was not covered by ORC, suggesting that ORC was not solely defining the NFR size. Nr-ACS were located inside much smaller NFRs with a weaker nucleosome depletion and with a symmetrically located ACS inside the NFR. The high A/T richness of nr-ACS known to physically exclude nucleosome could explain the formation of NFR at these positions. Finally, upstream and downstream nucleosomes lack the periodic positioning found at the ORC-ACS sites[116,117].

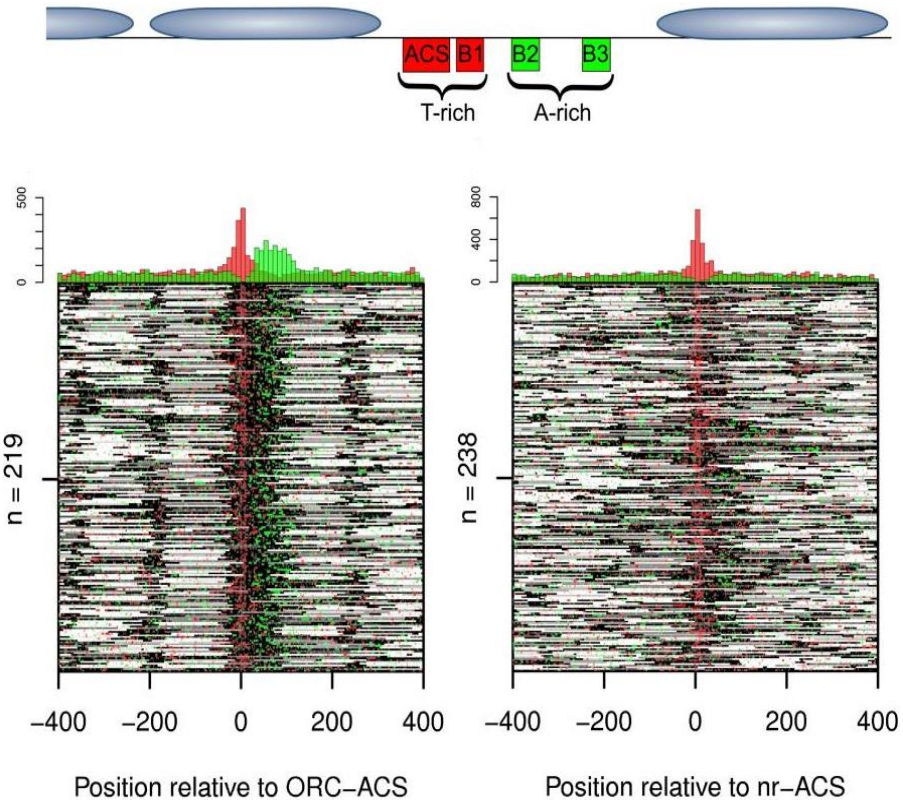


Figure 33

Organization of nucleosomes at budding yeast ARS. Heat-map of nucleosome occupancy at ORC occupied ARS (left) and non-replicative ARS (right). Black sequences represent non-nucleosome protected DNA sequences. Red and Green plots show the nucleotide bias at origins.

From Sarkies *et al*, (2010)

To investigate the role of primary DNA sequence in specifying the NFR and potentiating ORC binding, the distribution of *in vitro* assembled nucleosomes was analysed on both ORC-ACS and nr-ACS. Reconstitution on ORC-ACS, in the absence of ORC or any other *trans*-acting factors, recapitulates an imprecise similar NFR asymmetrically positioned over the ACS underlining the critical role of the sequence on nucleosome positioning. Similarly, nr-ACS also recapitulates *in vitro* the organisation found *in vivo*. Comparison of sequence composition associated with ORC-ACS and nr-ACS revealed that the presence A-rich islands located downstream of only ORC-ACS. The authors proposed that these A-rich elements maintain the large asymmetric NFR and thereby facilitate ORC binding. Finally, the periodic positioning of adjacent **nucleosomes requires ORC binding for a precise nucleosome positioning**.

To address whether the size of the NFR might increase when pre-RCs are loaded, cells have been synchronised in G1 (pre-RCs present) and in G2 (No pre-RC). S. *cerevisiae* cells have been synchronised in G2/M using nocodazole and at the end of G1 using α factor. In agreement with a role of the DNA sequence in NFR formation, NFR could be observed at ARS in G2 cells but, NFR during G1, depending on the origins, expanded due to an upstream or downstream nucleosome movement within 26% and 33% of NFR respectively. However, no nucleosome movement could be observed in both directions. Moreover, 41% of NFR found at origins were not affected[118]. These nucleosome movements were specific to replication origins and pre-RC formation since mutation of CDC6 did not induce any NFR expansion. Interestingly, most efficient origins were found to be bound by ORC during G2/M and showed an NFR expansion whereas inefficient or dormant origins were only bound by ORC during G1 and exhibited no NFR change[118]. Footprint of ORC and ChIP-Seq of MCM2-7 revealed a link between the orientation of the NFR expansion and pre-RC footprint. An expansion found upstream of the NFR expansion wass linked with an upstream pre-RC deposition. Coupled with MCM ChIP-Seq, it has been observed that the MCM once loaded were directly adjacent to the nucleosome[118].

Another study used ChIP-seq to analyse nucleosome occupancy at replication origins. Three classes of origins were defined as having high, medium and low nucleosome occupancy[119]. Nucleosome occupancy was commonly decreased during the G2/M to G1 transition and relied on the pre-RC assembly. Most efficient origins with low nucleosome occupancy tended to have a greater nucleosome loss during the G2/M to G1 transition and *vice-versa*. Interestingly, origins with low nucleosome occupancy were early firing origin, have a high efficiency and a high ORC binding profile. On the contrary, origins with high nucleosome occupancy were mostly late firing origin and have a low efficiency. Altering the nucleosome occupancy affects the origin usage and in part affects the replication timing program with 24% of late replicating origin becoming early replicating [119].

c. Nucleosomes and DNA replication in metazoan.

Compare to the yeast S. *cerevisiae*, only a limited number of genomic studies have been directed in metazoan.

As soon as the OGRE sequence had been identified and origins were mapped at a genome-scale level[68], nucleosome occupancy at origins had been investigated in mouse embryonic stem cells. A first analysis based on prediction of nucleosome occupancy suggested that regions centred on OGREs had on average a high intrinsic nucleosome occupancy [68]. However, experimental investigation of nucleosomepositioning over OGRE sequences led to different results. Two different sets of MNase-seq were used to make the analysis at the initiation sites. It turned out that these sites are associated with a**, well positioned nucleosomes**[70] (Figure 34). Taking advantage of the two MNase data sets corresponding to mild and strong digestions, positioned nucleosomes found at initiation sites were described as labile since they disappeared upon deeper MNase digestion. Finally, by contrast to what was proposed in the first study, OGRE sequences potentially forming G4 structure overlapped with NFRs found upstream of initiation sites. Origins that did not exhibit an OGRE sequence did not have a NFR but still had a well-positioned nucleosome over the initiation site[70].

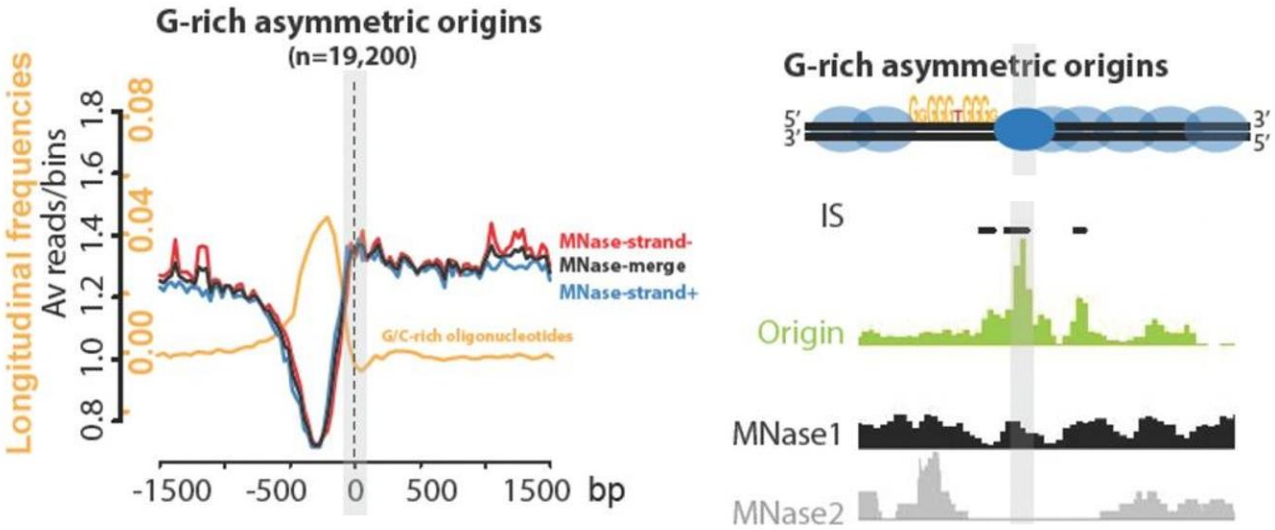


Figure 34

Nucleosome positioning at mouse replication origins. Above is shown a representation of nucleosome organization. Blue circles are nucleosomes. At the bottom are shown peak calling of SNS sample as well as two different MNase digestions.

From Cayrou *et al*, (2015)

Analysis of ChiP-seq data on **histone variants at the initiation site showed that the labile nucleosome was enriched in the variant H2A.Z**. However, 10 % of the H2A.Z sites only covered 8.6% of initiation sites[70]. Moreover, histone variant H3.3 was not found enriched at initiation sites. Altogether, these results suggest that histone variants are not involved in the observed "lability". Interestingly, the histone mark H3K64ac was observed at 86% of replication origins and displayed a high and sharp profile centred on the labile nucleosome at IS in all origin classes[70].

In Drosophila, ORC binding sites overlap with NDRs. Contrary to mice origins, the histone variant H3.3 was enriched in a close proximity of ORC binding sites even in non-promoter origins. The H2A.Z-like Drosophila histone variant H2Av, which is associated to +1 nucleosome of active promoters, was found associated with ORC but only on those found at active promoters[120].

At the dihydrofolate reductase locus in the mammal Chinese Hamster ovary cell line pre-RC seemed to co-localise with low nucleosome occupancy [121].

To investigate in more detail the interplay between nucleosome positioning and DNA replication origins, a precise approach using increasing levels of MNase digestion on native or cross-linked chromatin coupled with ChIP and followed by deep qPCR scanning has been carried on. Such technic allows the depiction of nucleosomes at a subset of origins. It was possible to distinguish stable nucleosomes that are not affected by deep MNase digestion from labile nucleosomes found only in low digestion MNase conditions or when cross-linking was applied. To precisely map replication initiation sites, very small nascent strands (600 bp) were purified. The active promoter studied did not exhibit any NFR, but instead presented a NDR identified as a labile nucleosome[122]. In studied origin subset, most efficient origins, associated with active promoter, presented a very well defined SNS peak localised at the +1 nucleosome. Origins with a lower efficiency exhibit SNS peaks dispersion but peaks still tended to be enriched at the positioned nucleosome independently of their stability. This result suggests a role of nucleosome positioning on origin activity. Non-promoter and low efficient origins did not present well positioned nucleosome and the SNS peaks were not well defined.

Altogether, these results indicate that replication initiation sites occur at positions of high-nucleosome occupancy and suggest that there is a correlation between stronger nucleosome positioning and higher origin efficiency.

To further investigate this relationship, the nucleosomes and SNS distribution have been studied at three different promoters in two cells lines having different levels of expression. Higher transcription affected the distribution of nucleosomes over the promoter together with SNS peaks. The co-localisation of SNS peaks with nucleosomes confirmed the previous conclusion.

Nucleosome positioning over the well-characterised laminB2 origin was analysed in HeLa synchronised cells. The nucleosome loss during S-phase right before replication origin activation suggests the removal of a labile nucleosome at the site of replication initiation.

Finally, ORC1 ChIP-seq peaks positioning was also investigated with respect to nucleosome stability. These peaks were found upstream of the TSS and associated with labile histone variants H2A.Z/H3.3 suggesting a link between the ORC binding sites and nucleosome lability[122]. Such observation is reminiscent of the observed higher histone turnover associated with ORC binding sites in Drosophila[123].

7. Research project

Despite all the efforts made to identify *cis*-regulatory elements involved in replication origins definition in vertebrates, no consensus sequence, playing a similar function as the ACS in *S. cerevisiae* origins, has been identified. However, G4 motifs have been found enriched within replication origins genome-wide[67,68,70], but most importantly, the formal role of G4 formation on origin activity was addressed genetically[72,73]. It became clear that G4 structure was essential for the activation of efficient replication origins. However, it was not sufficient to induce replication initiation as, in the chicken β^A-globin origin, the deletion of the 200bp long sequence, located just downstream of the pG4, disrupted the origin activity[72] (Figure 35).

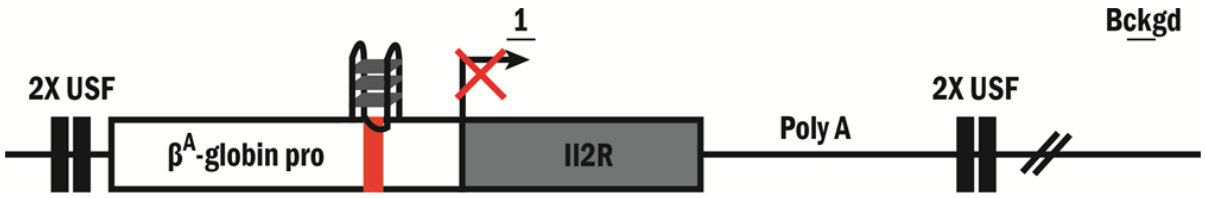


Figure 35

Chicken β^A-globin promoter associated with the IL2R gene flanked with USF binding sites. This construct is used as a model origin.

During my PhD I deleted and mutated several *cis*-elements known to be involved in transcription found in this sequence and investigated their effect on origin activity using SNS-purification and quantification. Besides, I also tested the origin behaviour regarding its capacity to advance the replication timing as a way to control SNS results and to address the potential role of these *cis*-regulatory elements on the replication timing control. Thanks to the results obtained with these deletions, I defined a minimal β^A-globin origin of only 90bp long allowing for the first time to define the minimal combination of key *cis*-elements for the formation of an active origin

I found that one G4 sequence coupled with another G-rich sequence, potentially forming a G-quadruplex and surrounded by a CAAT and a TATA-boxes were sufficient to induce formation of a functional origin. Moreover, I showed that the two pG4s have to be located on the same strand in order to initiate replication origin.

I also identified for the first time the role of the TATA-box on the replication-timing program. Further studies will be required to characterise molecular mechanisms involved in such timing program encoding.

To finish I analysed with the active minimal origin the causal relationship between formation of an active origin and nucleosome positioning. I observed a clear positioning of nucleosomes which is slightly modified when pre-RCs are loaded. The precise nucleosome organisation could not be observed on a non-active minimal origin that was modified only at the pG4#1, which was on the opposite strand. Altogether, this work opens new perspectives to decipher how the replication machinery is efficiently loaded and activated at replication origins in vertebrates.

Part2
Results

Two G4 motifs on the same strand associated with CCAAT and TATA-boxes shape a strong replication origin with well-positioned nucleosomes

Jérémy Poulet-Benedetti[1], Anne-Laure Valton[2], Caroline Brossas[1], Marc Laurent[1], Natalja Barinova[1], Aurore Champigny[1], and Marie-Noëlle Prioleau[1]*

[1] Université de Paris, CNRS, Institut Jacques Monod, F-75006, Paris, France

[2]Present address University of Massachusetts Medical School, Program in Systems Biology, Department of Biochemistry and Molecular Pharmacology, Worcester, MA 01605, USA

* Correspondence: marie-noelle.prioleau@ijm.fr

Abstract

DNA replication is under the control of a spatiotemporal program that ensures the accurate duplication of the genome. Many genome-wide studies in vertebrates identified promoters containing G-quadruplex motifs (pG4s) as being strongly associated with efficient replication origins. PG4 necessity has been proven genetically. However, it was shown that one pG4 was not sufficient to form an origin. We define here a minimal 90 bp model origin containing two pG4s located on the same strand and associated with a CCAAT and a TATA-box. The TATA-box contributes to its replication timing control. Distance between two cooperating pG4s is flexible, however, the transfer of one of them to the other strand abolishes origin function. Analysis of nucleosome organisation over the minimal origin reveals a very specific organisation previously found genome-wide in vertebrates. The origin is inside a nucleosome depleted region (NDR) and separated from the site of initiation by a strongly positioned nucleosome. In conclusion, our study provides a new paradigm for the genetic and nucleosome organisation of vertebrate origins.

Introduction

DNA replication is among the most critical steps that cells have to complete to ensure the proper transmission of their genetic material to daughter cells. Cells must accomplish, in a given time, the complete replication of their genome. To organise this complex process, eukaryotic cells are all characterised by a well-established replication timing program defining early, mid and late-replicating domains (Marchal et al., 2019). In addition, replication starts at sites called replication origins or inside initiation zones defining a spatial program that is intimately connected with the timing program (Ganier et al., 2019; Prioleau and MacAlpine, 2016). To fully understand how replication is regulated in eukaryotes, it is important to integrate these two parameters. Replication origins are defined by a primary sequence and a local chromatin structure that are recognized by essential initiator proteins. In the budding yeast *S.cerevisiae*, replication origins are called Autonomous Replication Sequence (ARS) containing an 11-bp long AT-rich consensus sequence (ACS) coupled with B1, and most of the time with B2 and B3 sequences. ARS are recognised, throughout the cell cycle, by the Origin Recognition Complex (ORC) proteins. During G1 in conjunction with Cdc6 and Cdt1, ORC recruits the Mcm2-7 helicase to initiation sites, leading to the formation of the pre-replicative complex (pre-RC), a process called licensing. Licensed origins are then progressively activated during S-phase according to the temporal program. A pioneering experiment at the ARS1 replication origin demonstrated that artificially positioning a nucleosome over the ACS impaired origin function, suggesting that nucleosomes govern ORC accessibility to the ACS (Simpson, 1990). Later on, this hypothesis was confirmed by the observation that an important feature of replication origins is a well-established nucleosome depleted region (NDR) flanked by nucleosomes asymmetrically positioned around the ACS (Berbenetz et al., 2010; Eaton et al., 2010). This control contributes to the relative small number (~220) of ORC-binding sites in the genome despite the presence of ~10 000 high-quality ACS motifs. It is also proposed that positioning of the flanking nucleosomes may play an active role in pre-RC formation as expanding the native NDR at ARS1 significantly limits pre-RC assembly (Lipford and Bell, 2001).

In vertebrates, identification of replication origins by several genome-wide approaches revealed a strong association of efficient origins with CpG islands (CGI) and promoters. In mouse and human, an origin G-rich repeated element (OGRE), potentially forming a G-quadruplex (pG4), a DNA secondary structure, was found enriched at replication origins (Besnard et al., 2012; Cayrou et al., 2012). PG4s are dynamic structures that have been firstly defined and computationally detected as four tracts of at least three guanines (G_3-N_{1-7}-G_3-N_{1-7}-G_3-N_{1-7}-G_3). Four guanines can interact through Hoogsten hydrogen bonds to form a G4 quartet which, when stacked on top of each other, form a G-quadruplex (G4). However, recently a high-resolution sequencing-based method (G4-seq) identified

~700 000 G4s inside the human genome, among which ~450 000 were not canonical G4 but contains either long loops or bulged structures (Chambers et al., 2015). Study using two model origins of replication in DT-40 chicken cell line demonstrated that pG4s are essential cis-elements for origin activity, thus confirming their importance in controlling replication initiation (Valton et al., 2014). This observation was also recently shown on mouse origins (Prorok et al., 2019). However, genome-wide studies and genetic studies showed that one pG4 is not sufficient to define an active origin underlying the complexity of their organisation in vertebrates. Chromatin organisation of replication origins has also been addressed in vertebrates. Surprisingly, sites of initiation mapped with the short nascent strand (SNS) method are defined by well positioned nucleosomes having properties of labile nucleosomes (Cayrou et al., 2015; Lombraña et al., 2013). Moreover, a nucleosome depleted region (NDR) was found over the OGRE element located upstream of the start site of active origins. Another key difference from yeast is that in human cells Orc1 is degraded in S-phase and has to be reloaded during early mitosis to assemble new pre-RCs during the following G1 phase (Kara et al., 2015). This points out that although the proteins forming the initiator are well conserved from yeast to human, the mechanisms that delineate preferential sites of initiation are different in terms of primary sequence and chromatin organisation and at least in the kinetics of pre-RCs loading during the cell cycle. To date, much information has been collected on the formation and activation of replication origins from *in vivo* and *in vitro* studies in *S.cerevisiae*. The field has reached a deep understanding of replication regulation in this model system. New observations made in vertebrates suggest that some concepts needed to be re-investigated in order to fully characterised molecular mechanisms shared or not with yeast. One key step is to delineate the minimal combination of critical *cis*-elements that can very efficiently define an origin of replication. In this study, we define a 90 bp very efficient origin containing two pG4s on the same strand with one being flanked by a CCAAT and a TATA-box. We analyse in details how nucleosomes are organised on this origin before and after pre-RCs loading. This study allows us to propose new models for origin activation in vertebrates.

Results

Transcription *cis*-regulatory elements are not essential for replication origin activity

Genome-wide mapping of replication origins in vertebrates has put forward several features associated with strong initiation sites including the presence of a transcription start site (TSS) and the association with G-quadruplex motifs (pG4). Genetic studies on model origins in chicken and mouse demonstrated the essential role of pG4 on origin activity. Moreover, the precise dissection of *cis*-elements composing the chicken β^A-globin promoter revealed that, although essential, one pG4 was not sufficient to trigger replication initiation (Valton et al., 2014). The 1.1 kb long β^A-globin promoter

associated with the Il2R reporter gene fused to a poly A sequence and flanked by two USF binding sites has the capacity to induce formation of a strong initiation site (Figure 1A)(Valton et al., 2014). Moreover, when ectopically inserted inside a mid-late replicating region devoid of strong replication origins, this construct has the capacity to advance replication timing (RT) locally in a cell population based assay through the presence of *cis*-regulatory elements involved in RT control (Supplementary Figure 1)(Hassan-Zadeh et al., 2012). Since only an efficient origin, which fires in most cells, can induce a significant RT shift, we concluded that this model origin is highly efficient. In agreement with this hypothesis, the short nascent strand (SNS) relative enrichment assay showed that the ectopic β^A-globin origin exhibited a relative enrichment similar to the endogenous β^A-globin functional origin (106% obtained with primer pair 1, grey bars, Figure 1B). SNS enrichment is normalised according to the endogenous ρ-globin origin that is artificially set to 100%; the enrichment found in a background region is quantified using a primer pair located 5kb downstream of the ectopic β^A-globin origin. Our previous study demonstrated the essential role in replication origin activity of one pG4 as well as a 245 bp long sequence located just downstream to it (pG4#1 and DNA sequence shown on Figure 1A, Supplementary Figure 2). Indeed, origin activity was drastically impaired by either pG4#1 or the 245 bp module deletion (Valton et al., 2014). To delineate new essential *cis*-motifs that could be common to metazoan replication origins embedded inside promoters, we focused on *cis*-regulatory elements found inside the 245 bp long module. This region contains erythroid-specific factors binding sites (CACC-box and NF-E4 binding site) as well as CCAAT and TATA-boxes which are binding sites for ubiquitous transcription factors NF-Y and TBP respectively (Figure 1A). Deletion of either the CCAAT or the TATA-box led to only a limited decrease of SNS relative enrichment (66% and 67% respectively with primer pair 1, light green and orange bars, Figure 1B). To investigate a potential cooperation between the CCAAT and TATA-boxes on origin function, we deleted both sequences. Surprisingly, the β^A-globin Δ CCAAT+TATA double mutant gave a SNS relative enrichment similar to the single CCAAT or TATA-box deletions (74% with primer pair 1, yellow bars, Figure 1B). These results suggested a limited but detectable implication of the CCAAT and TATA-boxes on replication origin activity.

A second, potentially forming G4 is essential for origin activity

To pursue our molecular dissection of *cis*-regulatory elements involved in the β^A-globin origin activity, we addressed the role of the CACC-box together with a pG4 overlapping this box. Indeed, a pG4 (named #2 and underlined, Figure 1A) is found 30 bp downstream of the essential pG4#1. pG4#2 is located on the opposite strand compared to pG4#1 and does not display the canonical sequence (G_3-$N_{1\text{-}7}$-G_3-$N_{1\text{-}7}$-G_3-$N_{1\text{-}7}$-G_3) initially used to search for pG4 genome-wide. However, it can form a pG4 either with a bulge or with a long loop (12 b). A third pG4 (pG4#3) is located between the CCAAT and the

TATA-boxes, on the same strand as pG4#1. However, pG4#3 does not present a canonical sequence but potentially form a G4 with a bulge. To investigate the role of the CACC-box together with pG4#2, we deleted the CACC-box as well as 12 bp downstream. This deletion (ΔCACC) led to a small increase of SNS relative enrichment suggesting that the origin is slightly more efficient (112% with primer pair 1, blue bars, Figure 1C). Deletion of the region from the beginning of the CCAAT-box to the end of the TATA-box (ΔCCAAT to TATA) induced a drastic drop in SNS relative enrichment (23 % with primer pair 1, dark green bars, Figure 1C). This region contains pG4#3 and its deletion almost recapitulated the SNS relative enrichment found when the entire 245 bp module was deleted, underlining the essential role of this sequence. To further characterise the essential role of pG4#3 on the origin activity, we replaced three guanines by three adenines. This mutant was expected to alter the formation of a G4 at pG4#3 (stars, Figure 1A). Those mutations, coupled with the CCAAT and TATA-box deletions (ΔCCAAT+TATA mpG4), gave a similar SNS enrichment as the complete CCAAT to TATA-box deletion (24% with primer pair 1, red bars, Figure 1C). Altogether, these results suggest that pG4#3 is critical for the origin activity. However, we could not precisely discriminate whether the G-richness or the formation of a G4 is the key regulatory element.

The TATA-box plays a role in RT control

We have previously developed a timing shift assay aimed at confirming the SNS relative enrichment method (Valton et al., 2014). In this assay, the ectopic replication origin activity is monitored by its capacity to locally advance the RT. This approach provides not only the information that the origin is functional but also that it carries its own timing information. We previously showed that the RT shift depends on both an active origin and the presence of two USF binding sites flanking the construct. We quantitatively measure the RT shift by calculating the -ΔL+ΔE number (Supplementary Figure 3). The more positive the -ΔL+ΔE value is, the stronger the RT shift is. From previous studies, we have observed that inefficient origins (SNS relative enrichment around 10%) are mostly replicated by the incoming forks from upstream and downstream origins and exhibit a -ΔL+ΔE median value of 6.5 (Inefficient β^A-globin origins, Figure 2A). Active β^A-globin origins having high SNS relative enrichment (around 100%) exhibit a significant RT shift with a -ΔL+ΔE median of 20.2 corresponding to a RT shift from mid-late to mid-early (Active β^A-globin origins, Figure 2). The CCAAT-box deletion mutant also induced a RT shift with a -ΔL+ΔE median value of 12.75 (β^A globin ΔCCAAT origin, Figure 2A and Supplementary Figure 4A). Surprisingly, deletion of the TATA-box induced a significant RT delay with a negative -ΔL+ΔE value of - 11.3 (β^A globin ΔTATA origin, Figure 2A and Supplementary Figure 5). This delay was not correlated with the SNS relative enrichment as it was the same for ΔCCAAT and ΔTATA-box mutants (Figure 1B). Similarly, a RT delay was observed with the CCAAT+TATA-boxes deletion mutant (-ΔL+ΔE median value

of -12.8, Figure 2A and Supplementary Figure 4B), suggesting the prevailing role of the TATA-box on RT control. The ΔCACC β^A-globin origin, that exhibited a high SNS relative enrichment, could also shift the RT (-ΔL+ΔE median of 19.1, Figure 2B and Supplementary Figure 6A). Finally, the β^A-globin origin mutants ΔCCAAT to TATA and ΔCAAT+TATA mpG4, leading to a drastic decrease in origin efficiency, did not induce any observable RT shift with a -ΔL+ΔE median value of 3.6 and 6.15 respectively (Figure 2B and Supplementary Figures 6B and 7). In these two last mutants, deletion of the TATA-box in a context of an inactive origin did not delay the RT in contrast with TATA-box deleted mutants that maintained a functional origin (β^A globin ΔTATA and ΔCCAAT+TATA origins). Altogether, these results confirmed our SNS relative enrichment data, but also indicated that inside this construct, the TATA-box is a key *cis*-element for RT control together with the already identified USF binding sites. It is quite puzzling to observe that its deletion in presence of an active origin led to a delay of RT as if the origin had some impact on the firing of flanking origins and/or replication fork progression inside the modified region. Further studies would be needed to investigate this point.

Definition of an active metazoan autonomous minimal replication origin

Our in-depth characterisation of essential *cis*-regulatory elements, constituting the β^A-globin origin, led us to consider whether these elements were sufficient to induce the formation of a functional and efficient origin. We thought that identification of such a minimal sequence would provide key information on the molecular events required to select and activate an origin. To answer this question, we designed a β^A-globin minimal origin deleted for all non-essential sequences. This 90 bp sequence only contained pG4#1, a small linker of 13 bp naturally present in the β^A-globin origin and the sequence from the CCAAT to TATA-box containing pG4#3 (Figure 3A). We investigated the origin activity using SNS quantification and the timing shift assay (Figures 3B and C). The β^A-globin minimal origin gave a SNS relative enrichment with primer pair 1 of 263% suggesting a 2.5-fold increase in origin efficiency compared to the complete β^A-globin origin (primer pair 1, Figure 3B). This high SNS enrichment could result from the primer pair movement toward pG4s found at the origin after the deletions of non-essential sequences (132 bp downstream of pG4#3 inside the minimal origin instead of 305 bp inside the full origin). To solve this issue, we created a new primer pair (named 1') located 317 bp downstream of pG4#3 inside the minimal origin, a position comparable to the primer 1 pair used on the complete β^A-globin origin. The new primer pair 1' gave a SNS enrichment of 125% that is similar to the one observed for the complete β^A-globin origin (Figure 3B). In agreement with a fully efficient origin, the timing shift obtained with the minimal origin gave a -ΔL+ΔE median value of 15.2 (Figure 3C and Supplementary Figure 8). Altogether, these results indicate that the minimal 90 bp β^A-globin origin

is active and efficient. Our data therefore showed that essential *cis*-elements proved to be sufficient to form a functional origin.

pG4#1 and #3 found inside the β^A-globin minimal origin have to be located on the same strand

We have previously shown that the orientation of pG4#1 determined the localisation of replication initiation 3' of the Gs tract and that its orientation could be changed with respect to the 245 bp module (Figure 4A) (Valton et al., 2014). Similarly, it was observed genome-wide in *Drosophila* that the orientation of pG4s defined the positioning of initiation sites (Comoglio et al., 2015). We therefore decided to test the impact of changing the strand of pG4#1 on the β^A-globin minimal origin. Surprisingly, in this new mutant no significant SNS enrichment could be detected either downstream or upstream of the pG4#1 (Primer pair 0 and 1 or 1', Figure 4B). This new intriguing result suggested that pG4#1 had to be localised on the same strand as pG4#3. It also questioned why pG4#1 could be inverted when associated with the 245 bp module. One hypothesis was that pG4#2 had also the capacity to cooperate with pG4#1 and thus to be part of a functional origin. To test the hypothesis, we took advantage of the previously constructed mutant of the β^A-globin origin deleted for pG4#1 (ΔpG4#1) and known to have no origin activity. We tried to rescue the origin activity by changing pG4#2 so that it would be on the same strand as pG4#3 (ΔpG4#1 pG4#2inv, Figure 4C). In agreement with our hypothesis, we could observe a diffused low but significant SNS relative enrichment over the IL2R sequence (around 30% at positions 1, 2 and 1', Figure 4C). To further characterise origin activity of this new construct, we tested the RT shift induced and observed a significant RT shift toward earlier replication (-ΔL+ΔE median value of 15.45, Figure 4C and Supplementary Figure 9). This last result confirmed that this origin is quite active and suggested that the result obtained with the SNS assay probably reflected a more diffuse site of replication initiation. Altogether, we observed that a functional origin depends on the presence of two pG4s oriented on the same strand. Moreover, we have found that all the combinations of pG4s found inside this model origin led to a functional origin. This last result showed that there is a high flexibility in the usage of pG4, although we do not have indication on whether some pG4s might be non-functional. Moreover, our results indicate that the distance between the two pG4s is not critical since several sizes of linker gave functional origins. However, we cannot exclude that a size limit between the two pG4s does exist.

The two essential pG4s found inside the Med14 promoter are not sufficient to form a functional origin

We have previously shown that *in situ* deletion of two pG4s oriented on the same strand inside the Med14 active promoter abolished its activity. This result is in line with our observation made on the β^A-globin origin showing that two pG4s cooperate to make a functional origin (Valton et al., 2014). We

also showed that deletion of one pG4 only partly affected the Med14 origin activity. At first glance, this last result seems to be in contradiction with the need for two pG4s to form a functional origin. Under careful look, it appeared that Med14 origin actually contains five pG4s on the same strand in a 1.1 kb long region. We previously choose to focus on the last two pG4s located next to the site of initiation (pG4#4 and 5)(Valton et al., 2014). In light of our new results, we can hypothesize that when pG4#4 or pG4#5 is deleted a third one, probably the closest one (pG4#3), can cooperate with the remaining pG4 (Figure 5A). Again, this is in agreement with the observation made on the β^A-globin origin in which all the combinations of two pG4s taken inside a set of three closely spaced pG4s can make an efficient origin. Firstly, we inserted a 529 bp region containing three pG4s # 3, 4 and 5 of the Med14 origin, at the same location as the one used to test the β^A-globin origin. After insertion, we observed a strong SNS relative enrichment (106%) in agreement with the formation of a strong origin at this position. We then inserted a minimal 100 bp long sequence containing only pG4s #4 and 5 (Figure 5B). Close examination of pG4 sequences suggest that they could form G4s with long loops. The Med14 minimal sequence was linked to the Il2R gene as well as the SV40 polyA sequence and flanked by two USF binding sites as for the β^A-globin minimal origin. SNS quantification showed that the Med14 minimal origin was not efficient (14% of SNS relative enrichment for both primer sets 1 and 1') (Figure 5B). To confirm this result we investigated the timing shift induced by the Med14 minimal origin and in agreement with the SNS analysis the construct did not exhibit any timing shift (–ΔL+ΔE median value of 4.4) (Figure 5B and Supplementary Figure 10). Taken together, these results suggest that the Med14 minimal origin, containing pG4 #4 and #5 that were proven to be necessary, could not support DNA replication. We here conclude that supplemental information needed to form a functional origin might be embedded into the sequence. Further studies are required to identify these *cis*-elements.

Nucleosomes are precisely positioned over the β^A-globin minimal origin

Now that we have defined a minimal and efficient origin, we decided to analyse how this 90 bp sequence might affect nucleosome organization. Indeed genome-wide studies in *S.cerevisiae* but also in human and mouse, suggest that a specific nucleosome organisation is found at replication origins. In vertebrates, two important observations were made. Firstly, initiation takes place at a well-positioned but labile nucleosome, secondly, strong origins containing G-rich motifs upstream of the initiation site display a Nucleosome depleted region (NDR) over this G-rich sequence. However, since most efficient origins are found at active TSS, it is difficult to uncouple the effect of transcription from the process of replication on nucleosome organisation. Our minimal β^A-globin origin does not drive transcription and therefore is a good model to explore the direct link between nucleosome

organisation and origin function (Supplementary Figure 11). We used MNase-Seq to map nucleosomes over our minimal origin. We took advantage of the inactive minimal origin in which pG4#1 was on the opposite strand as a negative control. To avoid contamination of nucleosome signal from the un-modified chromosome, we inserted the minimal and mutant origins on both chromosomes leading to homozygotes cell lines. Moreover, to have a dynamic vision of nucleosome positioning over our model origin, we synchronised cells at the G1/S transition (pre-RC bound) and in G2 phase (no pre-RC bound) (Supplementary Figures 12, 13, 14 and 15). We took advantage of the capacity to synchronize DT40 cells by elutriation. G1 elutriated cells were then treated for 3h with L-mimosine to block cells at the G1/S transition and to ensure a strong loading of pre-RC over replication origins whereas cells in G2 were directly processed after elutriation. Strikingly, we could observe five well-positioned nucleosomes over the IL2R reporter gene flanking a NDR localised on the β^A-globin minimal origin (Nucleosomes (Nuc) +1 to +5, Figure 6A). Nucleosomes +2 and +3 overlap and therefore cannot be found on the same chromosome. On chromosomes having either nucleosome +2 or +3, a space is available to fit a MCM double-hexamer (~70 bp). In G2, the nucleosome at position +3 is at least twice more abundant than the +2. Remarkably, at the G1/S transition when pre-RCs are loaded, the +3 nucleosome is less detectable and reach almost the same abundance as nucleosome +2. We therefore suggest that this change is linked to the loading of the pre-RCs at this position. We can envision that a fraction of chromosomes has a large NDR that allows the loading of several MCM double-hexamers. Upstream of the minimal origin (and the large NDR), nucleosomes organisation is less well defined but it is possible to delineate at least four discrete positions (Nuc -1 to -4, Figure 6A). Small differences can also be observed between G1/S and G2 especially at -2 and -3 nucleosomes. The -2 seemed to be better positioned during G1/S transition than at the G2 whereas it is the reverse for -3. Finally, the large NDR covered not only the 90 bp β^A-globin minimal origin but also about 100 bp upstream leaving also the 2XUSF binding sites uncovered. Altogether, we found an organisation that correspond to what has been described genome-wide for strong origins associated with a G-tract (Cayrou et al., 2015). This result validates our model origin as a great system to delineate basic components of what makes an efficient origin. To ensure that the nucleosome positioning is not imposed by the IL2R reporter gene independently of the replication origin, we analysed nucleosome positioning over the non-active β^A-globin minimal origin in which pG4#1 was on the opposite strand (Figure 6B). This origin contained exactly the same sequence as the β^A-globin minimal origin excepted at pG4#1. We expected to see a nucleosome positioning difference mostly linked to the origin activity and not to the sequence. Indeed, nucleosome landscape over the non-active β^A-globin minimal origin was extremely different from the one observed on the functional origin (Figure 6B). At first sight, the pattern is much more homogenous and especially no NDR could be observed over the 90 bp minimal non-functional origin. By contrast, we could observe a well-positioned nucleosome over this sequence suggesting that the presence of

two pG4s by itself is not sufficient to exclude nucleosome. Another interesting observation is that the +1 nucleosome is much less positioned, instead of two discrete positions named +2 and +3, one fuzzy nucleosome is found (+2/3), and finally +4 and +5 are displaced toward +2/3 limiting the space available in the non-functional minimal origin. Finally, the patterns observed at the end of S-phase and at the G1/S transition are quite similar (Figure 6B). Altogether, nucleosome positioning observed over and around the active β^A-globin minimal origin resembles the pattern found genome-wide on aggregate of origins containing a G-tract. Our study reveals that this precise organisation depends on the presence of two pG4s on the same strand and clearly demonstrates that one pG4 is not sufficient to organise a non-canonical pattern of nucleosomes. Moreover, since this minimal origin is not transcribed (Supplementary Figure 11), we can formally exclude a role of the transcriptional machinery in the establishment of this nucleosome pattern. It remains to understand how this sequence creates this organisation. The observation that the pattern is modified at the site of initiation when pre-RCs are formed suggests a role of components of the pre-RC in this establishment.

A strong nucleosome positioning is also observed over the entire β^A-globin origin

We showed a precise nucleosome organisation around the site of initiation of the β^A-globin minimal origin and we could link this organisation to origin activity. To strengthen our study, we decided to also analyse nucleosome organisation around the complete 1.1 kb long β^A-globin origin and several mutants in which origin efficiency or timing of activation were affected. Due to the presence of the endogenous β^A-globin origin sequence, we could not directly address the nucleosome positioning over the ectopic β^A-globin origin, as we would be contaminated by signal coming from the endogenous locus. To overcome this issue, we deleted the two endogenous β^A-globin origin sequences. Then we ectopically inserted the β^A-globin origin at the same position as previously studied and made MNase digestions on asynchronous cells. With the complete β^A-globin origin, we could observe a region of weak nucleosome positioning of approximately 400 bp directly upstream of pG4#1 (from 500 bp to 900 bp, Figure 7A). However, two discrete nucleosome positions could be identified (Nuc -1 and -2, Figure 7A). Upstream this region, a well-positioned nucleosome could be detected (Nuc -3). Interestingly, another highly positioned nucleosome was found just downstream of pG4#1 (Nuc +1, Figure 7A). This nucleosome covered pG4#2 and pG4#3 and had the space to exactly fit between pG4#1 and the TATA-box. This result is again in opposition to the idea that pG4s tend to exclude nucleosomes. Downstream the well-positioned +1 nucleosome (and the TATA-box), the signal could be interpreted as a fuzzy nucleosome or two populations of chromosomes with two different overlapping well-positioned nucleosomes as observed on the minimal origin. Here, we observed that only pG4#1 is located at the border of the nucleosome, whereas in the minimal origin, the two pG4s are located in a

NDR just upstream of a well-positioned nucleosome. This last result suggests that two pG4s located in a NDR is not a requirement for origin function.

We then asked how nucleosomes behaved when pG4#1 was deleted and the β^A-globin origin activity was strongly impaired. This mutant show a stronger enrichment over -1, -2 and -3 nucleosomes. Interestingly, the +1 nucleosome remained strongly positioned almost at the same location as the wild type origin. However, the stronger enrichment observed with the amplicon just upstream of the deleted pG4#1 suggested a slight movement of the nucleosome in the 5' direction (Figure 7B). We concluded that pG4#1 is not essential for nucleosome positioning but could act as a 5' barrier for nucleosome +1. The pattern downstream of the TATA-box and therefore at the site of initiation also seemed to be slightly affected.

To investigate the impact of replication initiation site on nucleosome positioning, we analysed the β^A-globin origin containing an inverted pG4#1. In this mutant, replication initiation site is shifted upstream of pG4#1 so that initiation remained 3' of the G tract (Figure 7C). The only clear difference in the nucleosome pattern was a slight movement of the +1 nucleosome in the 3' direction. Whether this movement would allow the binding of pre-RC nearby the inverted pG4#1 remained to be determined. However, it remains puzzling that nucleosome organisation remain close to the wild-type origin organisation, even though the replication initiation site has been drastically changed.

We further investigate nucleosome positioning on the non-active β^A-globin origin deleted from the CCAAT to the TATA-box (ΔCCAAT to TATA). Nucleosome positioning was globally affected, the strongest effect being observed at +1 and +2 nucleosomes, which appeared to get closer to each other and therefore possibly prevented the formation of a pre-RC at this position (Figure 8A).

Finally, we investigated nucleosome organisation of mutants deleted for the TATA-box and therefore affected in their RT control but not in their origin function. TATA-box deletion also resulted in perturbation +1 and +2 nucleosomes positioning whereas -1 to -4 nucleosomes were not affected (Figure 8B). Although the strong positioning of the Nuc +1 next to pG4#1 remained, Nuc +2 moved in the 5' direction and therefore overlapped with the Nuc +1. This profile could correspond to the overlap of two populations of chromosomes in which either Nuc +1 or Nuc +2 was present. The mutant deleted for both the CCAAT and TATA-boxes displayed the same nucleosome pattern in agreement with a major role of the TATA in the nucleosome organisation of this region (Figure 8C). This result suggested a strong role of the TATA-box in delineating the 3' border of nucleosome +1. Further studies will be needed to decipher the nucleosome organisation of this mutant and to understand its link with RT.

Discussion

Identification of ARS elements in *S.cerevisiae* has allowed to pinpoint a consensus sequence involved in the firing of most origins. This discovery was essential in deciphering both *in vitro* and *in vivo* molecular events involved in the activation of origins. The lack of a clear consensus sequence in vertebrate origins questions the mechanisms governing duplication of their genomes. Indeed, the observation that a robust temporal program can be obtained at the single cell resolution, especially in early and late S-phases, suggests that the replication initiation is at some point well-controlled (Duriez et al., 2019; Takahashi et al., 2019). To progress in this understanding, genetic studies aimed at manipulating endogenous *cis*-elements started to define critical elements (Hassan-Zadeh et al., 2012; Prorok et al., 2019; Sima et al., 2019; Therizols et al., 2014). The fastidious manipulation of complex genomes has greatly impeded genetic approaches. However, genome-wide studies put forward the presence of G-rich elements capable of forming a G-quadruplex inside most efficient origins (Besnard et al., 2012; Cayrou et al., 2012; Langley et al., 2016; Picard et al., 2014; Sugimoto et al., 2018). The use of the genetically amenable DT40 model system allowed us to demonstrate the essentiality of pG4s in origin activity (Valton et al., 2014). Here, we define new important features our two previously used model origins. These new elements provide key information, which will help us to better understand how the replication machinery is loaded and activated in vertebrates.

CCAAT and TATA-boxes roles on DNA replication initiation

CCAAT and TATA-boxes are *cis*-elements associated with many promoters although the TATA-box is present in only ~30% of promoters, TBP (TATA binding protein) is recruited at every active promoter. The ubiquitous transcription factors NF-Y and TBP both bend DNA to facilitate the transcription machinery loading. A large number of efficient origins firing at the beginning of the S-phase are associated with promoters in metazoan suggesting a complex regulatory link between transcription and replication. We could not detect expression of the *IL2R* gene neither on the complete β^A-globin promoter/origin nor on the minimal origin, allowing us to uncouple these two processes. Interestingly, several studies have shown that induction of transcription or simply an increased rate of transcription was associated with a RT program change suggesting a direct role of active transcription in triggering early firing (Blin et al., 2019; Brueckner et al., 2020). In our study, we found that CCAAT and TATA-boxes have a weak effect on origin efficiency *per se*. However, the TATA-box had a strong impact on RT as its deletion led to a RT delay whereas the CCAAT-box deletion had no effect on RT. This timing delay is dependent upon the presence of an active origin and we have not yet explore what could be the mechanism of regulation. We did not test whether TBP was recruited to the TATA-box but the absence of transcription suggests that there is no, or very transient, TBP recruitment. TBP could act on

the firing factors recruitment either by opening the chromatin structure, applying topological constraints to the chromatin or by directly interacting with some of them. We analysed how nucleosome positioning was affected in the mutant deleted for the TATA-box and found a profound effect on the positioning of Nuc+2 toward the 5' position suggesting that the TATA-box acts as a barrier against nucleosome invasion to Nuc +1. Further studies will be needed to decipher whether this barrier activity depends only on the sequence or whether TBP binding is involved in this barrier and if so, what is its role in the RT definition.

A specific nucleosome organisation is linked to origin function

Molecular dissection of the β^A-globin origin led to the identification, after pG4#1, of a second essential G-rich sequence. G4-seq revealed that pG4s, present in the human genome, can fold in a wide range of structures, including many G4 containing a bulge. These results strongly suggest that pG4#3 can be structured *in vivo*. However, we do not have any direct evidence of the role of the pG4#3 formation in origin function and we cannot discriminate whether the G-richness or G4 formation is the important feature. From our series of deletions, we could define an efficient β^A-globin minimal origin of only 90 bp long containing pG4#1 and pG4#3 flanked by a CCAAT and a TATA-boxes based on local enrichment in SNS and RT advancement. Despite its high efficiency, this origin is not the site of active transcription, making it an ideal system to study potential links between nucleosome organisation and origin function. We found a very precise organisation where the entire origin reside inside a NDR flanked by well-positioned nucleosomes. This organisation is reminiscent of the NDR found at functional ARS elements in *S.cerevisiae*. However, by contrast to what was observed in yeast, the site of initiation is not inside the NDR but ~ 200 bp downstream in between two very well positioned nucleosomes (+1 and +4, Figure 6A). Although we did not map MCM binding sites, it is reasonable to extrapolate that initiation should take place where MCM double-hexamer are formed. In agreement with this hypothesis, we found that this region is characterized by two exclusive nucleosome positions (Nuc+2 or Nuc+3, Figure 6A), each of them leaving a space for MCM double hexamer loading (~70 bp). We also observed a decrease in the coverage of Nuc+3 at the G1/S transition when pre-RCs are fully loaded. Further studies aimed at mapping MCMs should decipher whether this is indeed associated with a global increase in MCM loading. Such organisation questions how key *cis*-elements found inside the core origin might act on the loading of MCMs located remotely. One can envision that ORC, described to preferentially bind single stranded DNA G4 structure, recognized pG4#1 or/and #3 and therefore sits next to the site of DNA entry and exit of Nuc+1. Alternatively, ORC might have a better affinity for a NDR sitting next to well-positioned nucleosome. Again, further studies to precisely map ORC binding should bring more information on how this complex is loaded onto our minimal origin. This hypothesis

would suggest that pG4 promotes pre-RC formation, which is in contradiction with a recent study suggesting that OGRE/G4 sequences are essential for origin firing but not for their licensing (Prorok et al., 2019). The observation that MTBP (a partner of Treslin, the analogue of the yeast Sld3 firing factor), acts as a dimer and can preferentially recognize G4s, suggests that the minimal origin might also contribute somehow to origin firing (Kumagai and Dunphy, 2017). We also analysed with less resolution the nucleosome organisation of the full 1.1 kb β^A-globin origin. As on the minimal origin, we could identify just upstream of the site of initiation a strongly positioned nucleosome (Nuc+1, Figure 7A). This nucleosome seems to fit perfectly between the end of pG4#1 and the TATA-box, the length is exactly 147 bp, the size of the DNA wrapped around a nucleosome. Moreover, the site of initiation is also characterized by two overlapping nucleosome positions. However, by contrast to the observation made on the minimal origin only pG4#1 is found in a linker region whereas pG4#2 and 3 are within Nuc+1. This result confirmed a role of pG4 in adjusting the border of a nucleosome, however it also shows that pG4 can also be embedded in a highly positioned nucleosome. Altogether, we found a common organisation of nucleosomes of both types of origins.

Two pG4s on the same strand is a strong signal for origin function

Our study on the two endogenous Med14 and β^A-globin origins strongly suggest that origin organisation is flexible. Indeed, we could invert or delete one of the several pG4s naturally present inside the origin and yet maintain a substantial origin function. A deep analysis of how pG4s are organised inside these two model origins brought our attention toward the hypothesis that origin function could only be maintained when two pG4s are found to be oriented on the same strand. The analysis of nucleosome organisation suggests that it results in the strong positioning of a nucleosome next to a pG4 whereas the other one could be covered or not by a nucleosome. This result suggests that one final important signal might be a well-positioned nucleosome sitting next to a pG4. As we did on the β^A-globin origin, we tried to construct a minimal origin with the two pG4s found inside the Med14 promoter but we did not succeed. This result might reflect the fact that some supplementary *cis*-elements are also required. Indeed, the β^A-globin minimal origin contains also a CCAAT and a TATA-box. We need to test whether removal of these elements would block the activity of the minimal origin to define whether the association of others *cis*-elements are necessary to make this β^A-globin minimal origin functional. Genome-wide distribution of pG4 is not random. We questioned, genome-wide, the relationship between the density of G triplets tracks around pG4s and the activity of origins in DT40 cells. Firstly, we considered the G4s detected in our Ori-Seq dataset (Picard et al., 2014). We counted the number of G triplets in regions spanning 100 bp upstream and downstream of their pG4s

(Supplementary Figure 16). Then, we plotted the average origins activity according to the average number of G triplets in their associated pG4s. The trend is clear: the more pG4s have G triplets around, the more active origins are. This trend is less pronounced for origins associated with CGI (30% of chicken origins), indicating that for these origins, the contribution of the G triplets context is less important. This could be due to the contribution of binding sites for transcription factors. Then, we consider non Ori-pG4 as a negative control, we could see that the accumulation of reads is mildly affected by the density of G triplets around. Finally, we compared the G triplets density around Oris-pG4 and non-Oris pG4 and could see that Oris-pG4 have a higher density of G triplets around themselves than non-Oris-pG4 (Supplementary Figure 16C). Altogether, this analysis confirms genome-wide the observation on the contribution of two pG4s on origin function previously made on our model origins.

Materials & methods

Plasmid construction

To ensure homologous recombination in DT40 cells, replication origins have to be associated with BLS resistance gene, 5' and 3' arms, which are 2,5kb long homologous sequences to the insertion site (chr1:72,565,520 bp, galGal5), also present in the construct as described in Hassan-Zadeh *et al*, 2012. Transfection plasmids were constructed with the multisite Gateway Pro kit (Thermo Fischer Scientific #12537100). We used four entry vectors to generate the new β^A-globin +*IL2R* +β-actin +*BlsR* construct inserted at the desired mid-late genomic site. Two entry vectors containing the 5' and 3' target arms for specific insertion, one entry vector (β-actin*+BlsR*) containing the β-actin promoter linked with the *blasticidin* resistance gene (*BlsR*), flanked by *loxP* sites and one entry vector pDONR221 P5-P4 containing the β^A-globin fused to the *IL2R* gene and SV40 PolyA sequence, flanked with two USF binding sites. The corresponding final vector was generated by recombining compatible *att* sites between the entry vectors, with LR clonase. For electroporation, the final vector was linearized with *ScaI* (NEB #R3122S).

Origins mutagenesis were made using overlapping primers replicating the entire plasmids, using the Herculase II fusion DNA polymerase (Agilent) according to manufacturer recommendation. Primers were designed with around 15bp of overlapping sequence to ensure proper circularization with the In-Fusion HD cloning plus kit (Takara #638909), Mach1 competent cells (fisher scientific C862003) were used for plasmid cloning.

Cell culture condition and transfection

DT40 cells are grown in RPMI 1640 medium supplemented with Glutamax (Thermo Fisher Scientific #61870010), containing 10% FBS, 1% chicken serum, 0.1mM β-mercaptoethanol, 200U/mL penicillin and 200µg/mL streptomycin and 1,75 µg/mL of amphotericin B. Cells are grown at 37°C and under an atmosphere containing 5% CO_2. Cells were electroporated as previously described in Hassan-Zadeh *et al*, 2012. To address proper homologous recombination of DT40 transfected cells, genomic DNA was extracted from 400µl of 10 days cultured unique clones after digestion of 1h at 37°C using lysis buffer (10mM Tris pH8, 25mM NaCl, 1mM EDTA, 200µg/mL Proteinase K). Proteinase K was inactivated with a 10min incubation at 95°C. The genotyping PCR was performed using primer pairs located on the construct and on the genomic DNA (outside of homologous arms) with the Herculase II Fusion DNA polymerase system.

The *blasticidin* resistance gene was excised from positives clones using the Cre-LoxP system. DT40 cells constitutively expresses a tightly regulated Cre recombinase fused to a mutated estrogen receptor (Mer) that localised the protein in the cytoplasm. Addition of 4-hydroxytamoxifen (Sigma Aldrich #T176) result in Cre import to the nucleus, that induce the efficient excision of genomic regions flanked by two recombination signals (*loxP* sites) inserted in the same direction. The treatment of 3 x 10^5 cells with 5µM hydroxyl-tamoxifen for 24h results in the excision of the *BlsR* gene. The correct excision was controlled using primer pairs localised on the 3' arm and in the SV40 polyA sequence. For each clones, copy number of the constructs are quantified by qPCR (Hassan-Zadeh et al., 2012).

Homozygous insertions were made in a two step process. Cells were first heterozygously modified as described earlier, screened and then the *BlsR* gene was excised. Once cell lines were established, the second homologous recombination was made on the other chromosome. Cells were then screened depending on their proper insertion (primer located inside the *BlsR* gene and outside the 3' arm). In parallel, to ensure the first construct insertion was still present, the *BlsR* excision PCR was realised.

Cells synchronisation in G1/S was performed using 160 µg/ml of L-mimosine for 3h (Sigma Aldrich M0253).

SNS purification

SNS purification was performed as previously described in Valton *et al*, 2014 with some slight modifications. Fresh cultured cells were used for total genomic DNA extraction and the Polynucleotide kinase (New England biolabs) concentration was adjusted to 100U and incubated for 30min at 37°C.

Proteinase K (thermo Fischer #EO0491) digestion was realized at final concentration of 625µg/ml for 30min at 50°C.

Replication timing assay

Replication assay was carried as previously described in Hassan-Zadeh *et al.* 2012 (see Supplementary Figure 3). About 10^7 exponentially growing cells were pulse-labeled with 5-Bromo-2'-deoxyuridine (BrdU, Sigma-Aldrich #B9285) for 1 h and fixed in 75% of cold ethanol. Before cell sorting, cells were resuspended at a final concentration of 2.5 x 10^6 cells/mL in 0.1% IGEPAL in PBS (Sigma, #CA-630), the DNA was labelled with 50 µg/ml propidium iodide and RNA molecules were degraded using 0.5 mg/ml RNase A during a 30 minutes incubation period at room temperature. Cell sorting was realised at the ImagoSeine facility at Jacques Monod Institute with an INFLUX 500 cell sorter (Cytopeia, BD Biosciences). Four fractions of S-phase cells from early to late S phase, each containing 5x10^4 cells were collected and further treated for RT analyses. The collected cells were treated with lysis buffer (50 mM Tris pH 8.0; 10 mM EDTA pH 8.0; 300 mM NaCl; 0.5% SDS, 0.2 mg/ml of freshly added proteinase and 0.5 mg/ml of freshly added RNase A), incubated at 65°C for 2 h and stored at -20°C, in the dark. Genomic DNA was isolated from each sample by phenol-chloroform extraction and alcohol precipitation and sonicated four times for 30s each, at 30s intervals, in the high mode at 4°C in a Bioruptor water bath sonicator (Diagenode), to obtain fragments of 500 to 1000 bp in size. The sonicated DNA was denatured by incubation at 95°C for 5 minutes. We added monoclonal anti-BrdU antibody (BD Biosciences #347580) at a final concentration of 3.6 µg/ml in 1x IP buffer (10 mM Tris pH 8.0, 1 mM EDTA pH 8.0, 150 mM NaCl, 0.5% Triton X-100, and 7 mM NaOH). We used 30 µl or 50 µl of protein-G-coated magnetic beads (from Ademtech #4342 or Thermo Fisher Scientific #10004D, respectively) per sample to pull down the anti-BrdU antibody. Beads and BrdU-labeled nascent DNA were incubated for 2-3 hours at 4°C, on a rotating wheel. The beads were then washed once with 1x IP buffer, twice with wash buffer (20 mM M Tris pH 8.0, 2 mM EDTA pH 8.0, 250 mM NaCl, 0.25% Triton X-100) and then twice with 1x TE buffer pH 8.0. The DNA was eluted by incubating the beads at 37°C for 2 h in 250 µl 1x TE buffer pH 8.0, to which we added 1% SDS and 0.5 mg/ml proteinase K. DNA was purified by phenol-chloroform extraction and alcohol precipitation and resuspended in 50 µl TE.

MNase Digestion

We cross-linked 30 x 10^6 exponential growing cells by incubation for 5 minutes with 1% freshly prepared formaldehyde at room temperature (Thermo Fisher Scientific, #28908). Fixation was stopped by adding 0.125 M glycine-PBS for five minutes at room temperature. After three ice cold PBS washing, nuclei were extracted using lysis buffer + triton (10mM Tris-HCl, pH7.5, 10mM NaCl, 3mM $MgCl_2$, 0.2%

triton X-100, 0.5mM EGTA, 1mM DTT, 1x protease inhibitor cocktail (Sigma, #P8340)) for 5min on ice. Nuclei were then washed with lysis buffer minus Triton (10mM Tris-HCl, pH7.5, 10mM NaCl, 3mM $MgCl_2$, 0.,5mM EGTA, 1mM DTT, 1x protease inhibitor cocktail (Sigma, #P8340)) and resuspended in digestion buffer (10mM Tris-HCl, pH7.5, 10mM NaCl, 3mM $MgCl_2$, 1mM $CaCl_2$, 1x Protease inhibitors (Sigma, #P8340)). Nuclei were adjusted to 1,7 x 10^6 nucleus/mL and then digested using 2,5U, 10U, 40U and 160U of Micrococcal nuclease (MNase; Thermo Fischer Scientific #EN0181) for precisely 15min at 37°C with 2min pre-incubation. Reactions were stopped by adding (20mM EDTA pH 8, 4mM EGTA pH 8) and placed on ice. Samples were then treated with 80µg of RNase A for 30min at 37°C and with 100µg of proteinase K for an overnight incubation at 65°C. DNA was purified using phenol chloroform extraction. DNA sample was then subjected to size selection using Beckman Coulter SPRISelect beads (Beckman coulter #B23317) in agreement with the manufacturer instruction. 0.5X and 1.3X beads were used for 20ng/µL of digested DNA. The purified DNA was then quantified by real-time qPCR. High throughput sequencing was performed using digested DNA without SPRIselect size selection.

RNA extraction and reverse transcription

Total RNA were extracted from 5x10^6 cells with the Nucleospin RNA kit (Macherey Nagel, #740955). 20 µg of total RNA was then treated with 4 units of DNAse I (NEB, #M0303S) for 1 h at 37°C. The enzyme was inactivated by adding 5 mM EDTA and incubating the reaction mixture for 10 min at 75°C. The RNA was then purified by phenol-chloroform extraction and ethanol precipitation. Reverse transcription reactions (RT +) were then performed with 5 µg of RNA and random hexamers (NEB, S1330S), using the Superscript III Reverse Transcriptase (Thermo Fisher Scientific, #18080093) according to the manufacturer's instructions. Negative controls (RT -) were performed with the same procedure, but without the addition of reverse transcriptase. The comparison of RT + and RT - samples was used to validate DNAseI treatment and the complete digestion of the genomic DNA in the RNA samples.

Real-time PCR quantification of DNA

Real-time qPCRs were executed according to the MIQE guideline. The Techne Prime Pro48 apparatus and the QPCR-SYBR Green mix (Thermo Fisher Scientific, #AB1285B) were used for the real-time PCR quantification of BrdU-labeled nascent strands (NS), genomic DNA extracted from 4-hydroxytamoxifen-treated clonal cell lines, short nascent strands or cDNA. Each samples were quantified at least in duplicates. For all reactions on the Techne Prime Pro48, real-time PCR was performed under the following cycling conditions: initial denaturation at 95°C for 15 minutes, followed by 50 cycles of 95°C for 15 s, 61°C for 30 s, 72°C for 20 s, and fluorescence measurement. Following

PCR, a thermal melting profile was used for amplicon identification. MNase digestion protected DNA quantification on complete β^A-globin origin was quantified using the Roche Light Cycler 480 detection system with LightCycler 480 Sybr Green master mix (Roche Applied Science, # 04707516001). Samples were quantified in triplicates. We normalise our samples using two distinct amplicons located inside the condensed chromatin region of the endogenous *β-globin* locus. For each clonal cell line, the mean values obtained for the two condensed genomic regions for each digestion were arbitrarily set to 1 and used to normalize independent clones for each transgenic line separately and between all transgenic lines. The PTHLH insertion site and the 5kb downstream site were also used as controls characterizing the targeted locus. Real-time PCR was performed under the following cycling conditions: initial denaturation at 95°C for 5 minutes, followed by 50 cycles of 95°C for 10 s, 61°C for 20 s, 72°C for 20 s and fluorescence measurement. Following PCR, a thermal melting profile was used for amplicon identification. cDNA was quantified using primer located inside the *Il2R* gene and the *Med14* gene was used as reference.

Cells sorting by centrifugal elutriation

Elutriation was performed as previously described in (Duriez et al., 2019) with slight modifications. We used a Beckman Coulter Avanti® J-26 XP with a JE-5.0 rotor. Around 1.5 x 10^9 cells were resuspended into 200ml of elutriation buffer (PBS1X, 1% FBS and 1mM EDTA; filtered on 0.22µm). Cells are injected into an elutriation chamber of (40ml) subjected to a rotation of 2700 RPM, with a pump flow of 80ml/min. Once injected, cell population was incubated for 30min into the chamber for recovery and equilibration. Cells are sorted by size with decreasing centrifugal speed. We recover five fractions of cells (from F1 to F5) with the following parameters of centrifugal force / recovered volume: 2530RPM/400ml; 2380RPM/800ml; 2260RPM/800ml; 2130RPM/800ml; 1980 RPM/800ml. The Fraction 2 corresponds to G1 cells and the fraction 5 to G2/M cells. Cells were then cultured with classical medium culture at a concentration of 1 x 10^6 cells/ml. They were either directly treated with L-mimosine for 3h, or directly cross-linked and extracted for MNase digestion.

Sequencing library preparation

Sequencing libraries were prepared using NEBNext Ultra II DNA library prep Kit for Illumina (NEB #E7645S) in agreement to manufacturer instructions. For MNase libraries prep, samples were not subjected to size selection, but only cleaned-up for adaptor-ligated DNA using SPRISelect Reagent kit (Beckman coulter #B23317), to ensure a MNase digested small fragment overall recovery. Libraries were labelled with the NEB-Next Multiplex Oligos for Illumina set 3 and set 4. Libraries were prepared with a starting DNA input of 100ng. The mean size of the library molecules and the quality of the

libraries were determined on an Agilent Bio-analyser High Sensitivity DNA chip (Agilent technologies, #5067–4626) and with qPCR quantifications.

Sequencing

Sequencing was made at the GENOM'IC cochin institute facility (Paris) ona NextSeq 500 Illumina sequencer. Samples were sequenced with a High Output Flow Cell in paired-end using 200 M of reads for β^A-globin minimal origin and 100 M of reads for inverted G4 β^A-globin minimal origin.

References

Berbenetz, N.M., Nislow, C., and Brown, G.W. (2010). Diversity of eukaryotic DNA replication origins revealed by genome-wide analysis of chromatin structure. PLoS Genet. *6*, e1001092.

Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., Marin, J.-M., and Lemaitre, J.-M. (2012). Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nat. Struct. Mol. Biol. *19*, 837–844.

Blin, M., Le Tallec, B., Nähse, V., Schmidt, M., Brossas, C., Millot, G.A., Prioleau, M.-N., and Debatisse, M. (2019). Transcription-dependent regulation of replication dynamics modulates genome stability. Nat. Struct. Mol. Biol. *26*, 58–66.

Brueckner, L., Zhao, P.A., van Schaik, T., Leemans, C., Sima, J., Peric-Hupkes, D., Gilbert, D.M., and van Steensel, B. (2020). Local rewiring of genome-nuclear lamina interactions by transcription. EMBO J. *39*, e103159.

Cayrou, C., Coulombe, P., Puy, A., Rialle, S., Kaplan, N., Segal, E., and Méchali, M. (2012). New insights into replication origin characteristics in metazoans. Cell Cycle *11*, 658–667.

Cayrou, C., Ballester, B., Peiffer, I., Fenouil, R., Coulombe, P., Andrau, J.-C., van Helden, J., and Méchali, M. (2015). The chromatin environment shapes DNA replication origin organization and defines origin classes. Genome Res. *25*, 1873–1885.

Chambers, V.S., Marsico, G., Boutell, J.M., Di Antonio, M., Smith, G.P., and Balasubramanian, S. (2015). High-throughput sequencing of DNA G-quadruplex structures in the human genome. Nature Biotechnology *33*, 877–881.

Comoglio, F., Schlumpf, T., Schmid, V., Rohs, R., Beisel, C., and Paro, R. (2015). High-resolution profiling of Drosophila replication start sites reveals a DNA shape and chromatin signature of metazoan origins. Cell Rep *11*, 821–834.

Duriez, B., Chilaka, S., Bercher, J.-F., Hercul, E., and Prioleau, M.-N. (2019). Replication dynamics of individual loci in single living cells reveal changes in the degree of replication stochasticity through S phase. Nucleic Acids Res. *47*, 5155–5169.

Eaton, M.L., Galani, K., Kang, S., Bell, S.P., and MacAlpine, D.M. (2010). Conserved nucleosome positioning defines replication origins. Genes Dev. *24*, 748–753.

Ganier, O., Prorok, P., Akerman, I., and Méchali, M. (2019). Metazoan DNA replication origins. Curr. Opin. Cell Biol. *58*, 134–141.

Hassan-Zadeh, V., Chilaka, S., Cadoret, J.-C., Ma, M.K.-W., Boggetto, N., West, A.G., and Prioleau, M.-N. (2012). USF binding sequences from the HS4 insulator element impose early replication timing on a vertebrate replicator. PLoS Biol. *10*, e1001277.

Kara, N., Hossain, M., Prasanth, S.G., and Stillman, B. (2015). Orc1 Binding to Mitotic Chromosomes Precedes Spatial Patterning during G1 Phase and Assembly of the Origin Recognition Complex in Human Cells. J. Biol. Chem. *290*, 12355–12369.

Kumagai, A., and Dunphy, W.G. (2017). MTBP, the partner of Treslin, contains a novel DNA-binding domain that is essential for proper initiation of DNA replication. Mol. Biol. Cell *28*, 2998–3012.

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

Lipford, J.R., and Bell, S.P. (2001). Nucleosomes positioned by ORC facilitate the initiation of DNA replication. Mol. Cell *7*, 21–30.

Lombraña, R., Almeida, R., Revuelta, I., Madeira, S., Herranz, G., Saiz, N., Bastolla, U., and Gómez, M. (2013). High-resolution analysis of DNA synthesis start sites and nucleosome architecture at efficient mammalian replication origins. EMBO J. *32*, 2631–2644.

Marchal, C., Sima, J., and Gilbert, D.M. (2019). Control of DNA replication timing in the 3D genome. Nat. Rev. Mol. Cell Biol. *20*, 721–737.

Picard, F., Cadoret, J.-C., Audit, B., Arneodo, A., Alberti, A., Battail, C., Duret, L., and Prioleau, M.-N. (2014). The Spatiotemporal Program of DNA Replication Is Associated with Specific Combinations of Chromatin Marks in Human Cells. PLoS Genet *10*, e1004282.

Prioleau, M.-N., and MacAlpine, D.M. (2016). DNA replication origins-where do we begin? Genes Dev. *30*, 1683–1697.

Prorok, P., Artufel, M., Aze, A., Coulombe, P., Peiffer, I., Lacroix, L., Guédin, A., Mergny, J.-L., Damaschke, J., Schepers, A., et al. (2019). Involvement of G-quadruplex regions in mammalian replication origin activity. Nat Commun *10*, 3274.

Sima, J., Chakraborty, A., Dileep, V., Michalski, M., Klein, K.N., Holcomb, N.P., Turner, J.L., Paulsen, M.T., Rivera-Mulia, J.C., Trevilla-Garcia, C., et al. (2019). Identifying cis Elements for Spatiotemporal Control of Mammalian DNA Replication. Cell *176*, 816-830.e18.

Simpson, R.T. (1990). Nucleosome positioning can affect the function of a cis-acting DNA element in vivo. Nature *343*, 387–389.

Sugimoto, N., Maehara, K., Yoshida, K., Ohkawa, Y., and Fujita, M. (2018). Genome-wide analysis of the spatiotemporal regulation of firing and dormant replication origins in human cells. Nucleic Acids Res. *46*, 6683–6696.

Takahashi, S., Miura, H., Shibata, T., Nagao, K., Okumura, K., Ogata, M., Obuse, C., Takebayashi, S.-I., and Hiratani, I. (2019). Genome-wide stability of the DNA replication program in single mammalian cells. Nat. Genet. *51*, 529–540.

Therizols, P., Illingworth, R.S., Courilleau, C., Boyle, S., Wood, A.J., and Bickmore, W.A. (2014). Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. Science *346*, 1238–1242.

Valton, A.-L., Hassan-Zadeh, V., Lema, I., Boggetto, N., Alberti, P., Saintomé, C., Riou, J.-F., and Prioleau, M.-N. (2014). G4 motifs affect origin positioning and efficiency in two vertebrate replicators. EMBO J. *33*, 732–746.

Figure legends

Figure 1: Identification of essential *cis*-elements inside the β^A-globin origin

(A) Schematic representation of the β^A-globin origin (at scale) coupled with the *IL2R* gene, SV40 polyA sequence and flanked by USF binding sites. Black lines with numbers above represent the position of amplicons used for SNS quantifications. The sequence downstream of the pG4#1 is shown with different boxes and binding sites found. Coloured boxes represent the different elements and corresponding deletions made. The CACC-box deletion extends 12bp downstream of the blue box. The NF-E4 binding site is surrounded by a black rectangle. Potential G4 (pG4) are shown underlined. Stars are the mutated bases in the ΔCCAAT+TATA mpG4 mutant (G to A). (B and C) Relative SNS enrichments in clones with deletions of *cis*-regulatory elements associated with transcription inside the β^A-globin origin (B) or mutations inside the pG4#3 (C) are shown. Errors bars indicate the standard deviation for SNS enrichment in two independent clones and qPCR duplicates. For each clone, two amplicons at the initiation sites of the ectopic origin (1 and 2), one located 5 kb away of the site of integration to quantify the background signal (Bckg), and one amplicon specific of the endogenous β^A-globin origin (Endogenous β^A-globin) were used together with the endogenous ρ-origin to quantify SNS abundance.

Figure 2: RT shift assay reveals an un-expected role of the TATA box on timing

(A-B) Distribution of the RT shift of the different β^A-globin mutant origins. (A) RT shifts from either inefficient β^A-globin origins having a relative SNS enrichment around 10% or active β^A-globin origins having a relative SNS enrichment around 100% were collected from previous studies. RT shifts obtained with the new mutants described in Figure 1 were shown. Statistical analysis was performed with Wilcoxon non-parametric two-tailed tests for assays made with more than two clones (***p-value <0.001). (B) Distribution of RT shifts obtained on mutants affecting pG4#2 or #3 were represented. Dotted lines represent empirically set boundaries of RT shift that are not significant. Rectangle edges correspond to the 0.25 and 0.75 quartiles, the thick red lines represent the median and the whiskers extend to the smallest and largest -ΔL+ΔE values.

Figure 3: A 90bp β^A-globin minimal origin drives efficient DNA replication initiation

(A) Scheme of the β^A-globin minimal origin coupled with the *IL2R* gene, SV40 PolyA sequence and USF binding sites. Downstream the complete sequence of the minimal origin with the pG4s underlined and *cis*-regulatory elements delineated by coloured boxes are shown. Black lines with numbers above represent amplicons used for SNS qPCR analyses. (B) Relative SNS enrichments obtained with the β^A-globin minimal origin. Errors bars indicate the standard deviation for SNS enrichment in two independent clones and qPCR duplicates. (C) Distribution of RT shift obtained with inefficient and efficient origins previously described in Figure 2 with new RT shift values for the β^A-globin minimal origin. Rectangle edges correspond to the 0.25 and 0.75 quartiles, the thick red lines represent the median and the whiskers extend to the smallest and largest -ΔL+ΔE values. Statistical analysis was performed with Wilcoxon nonparametric two-tailed tests (ns, not significant; *** = p-value <0.001). Dotted lines represent empirically set boundaries of significant RT shift.

Figure 4: pG4 orientation impacts on the activity of the minimal origin

(A) Schematic representation of the impact of pG4#1 orientation on the localisation of replication initiation. The white box represents the 245 bp cooperating module identified in Valton *et al* (2014). Red molecules corresponds to leading strand, blue molecules to Okazaki's fragments, and green segments to the RNA primers. (B) Scheme of the β^A-globin minimal origin containing pG4#1 on the opposite strand. Downstream, the complete sequence is shown with pG4 underlined and *cis*-regulatory elements defined by coloured boxes. Black lines with numbers above are amplicons used for SNS qPCR analyses. Relative SNS enrichments are represented below. Errors bars indicate the standard deviation for SNS enrichment in two independent clones and qPCR duplicates. (C) Schematic representation of β^A-globin origin with the ΔpG4#1 and the pG4#2 on the opposite strand, fused with *IL2R* gene, SV40 PolyA sequence and USF binding sites. The nucleotide sequence from ΔpG4#1 to *Il2R* gene is shown below. The CACC-box is disrupted here due to pG4#2 being on the opposite strand. We enlarge the inverted sequence of pG4#2 to 50 bp (bases written in bold case) instead of only 36bp to prevent the formation of a new pG4 on the other strand, due to its natural G-richness. Relative SNS enrichments are shown below on the left. Errors bars indicate the standard deviation for SNS enrichment in two independent clones and qPCR duplicates. The distribution of RT shifts are shown on the right. Rectangle edges correspond to the 0.25 and 0.75 quartiles, the thick red lines represent the median and the whiskers extend to the smallest and largest -ΔL+ΔE values.. Statistical analysis was performed with Wilcoxon nonparametric two-tailed tests (** 0.001 < p-value < 0.01). Dotted lines represent empirically set boundaries of significant RT shift.

Figure 5: Two pG4s found inside the Med 14 origin are not sufficient to form an active origin

(A) Scheme of the Med14 origin inserted at the ectopic position with the partial sequence indicated below. PG4s are underlined and indicated by red boxes. The brown box represents a remaining sequence generated during the cloning procedure and still associated with the construct. Black lines with numbers are amplicons used for SNS qPCR analyses. Relative SNS enrichments are shown below. Errors bars indicate the standard deviation for SNS enrichment in two independent clones and qPCR duplicates. (B) Scheme of the Med14 minimal origin coupled with the *IL2R* gene, SV40 PolyA sequence and USF binding sites. The complete sequence of the minimal origin is given below with the pG4s underlined and delineated by red boxes. Black lines with numbers above represent amplicons used for SNS qPCR analyses. Relative SNS enrichments are shown below on the left. Errors bars indicate the standard deviation for SNS enrichment in two independent clones and qPCR duplicates. The distribution of RT shift is presented on the right. Rectangle edges correspond to the 0.25 and 0.75 quartiles, the thick red lines represent the median and the whiskers extend to the smallest and largest -ΔL+ΔE values. Dotted lines represent empirically set boundaries of significant RT shift.

Figure 6: The β^A-globin minimal origin induces a precise nucleosome organisation

(A) Nucleosomes map over the β^A-globin minimal origin. Numbers represent nucleosomes positioned according to the minimal origin, from +1 to +5 downstream and from -1 to -4 upstream. The graph represents mono-nucleosomes accumulation profiles along the sequence obtained in cells synchronised at the G1/S transition (red line) or in G2 (blue). Bottom schemes represent the deduced nucleosome positioning over the construct depending on the cell cycle. (B) Nucleosomes map over the inverted pG4#1 β^A-globin minimal origin. Numbers represent nucleosomes positioned downstream to the minimal origin from +1 to +5. G1/S transition synchronised cells are in red and G2 synchronised cells in blue.

Figure 7: Nucleosome positioning over the β^A-globin origin also reveals a specific pattern

(A-C) Schemes of the non-modified β^A-globin origin (A), the β^A-globin ΔpG4#1 origin (B) and the β^A-globin inverted pG4#1 origin (C) coupled to the *IL2R* gene (partly shown) and associated with two upstream USF binding sites. Upstream, relative enrichments of DNA found inside mono-nucleosomes purified after MNase digestion of the corresponding clone deleted for endogenous $\Delta\beta^A$-globin are shown. Numbers (from +1 to -4) represent nucleosomes relative to the pG4#1.

Figure 8: The TATA-box is involved in nucleosome positioning

(A-C) Schemes of the β^A-globin ΔCCAAT to TATA origin (A), the β^A-globin ΔTATA origin (B) and the β^A-globin ΔCCAAT+TATA origin (C) coupled to the *IL2R* gene (partly shown) and associated with two upstream USF binding sites. Upstream, relative enrichments of DNA found inside mono-nucleosomes purified after MNase digestion of the corresponding clone deleted for endogenous $\Delta\beta^A$-globin are shown and compared with the relative enrichments of DNA corresponding to the non-modified β^A-globin origin represented in figure 7A. Numbers (from +1 to -4) represent nucleosomes relative to the pG4#1.

Supplementary figure legends

Supplementary Figure 1: Properties of the insertion site targeted in this study

(A) UCSC genome browser visualization of the mid-late insertion site of chromosome 1 (galGal5). Single reads form SNS aligned determined in (Massip *et al*, 2019) are reported. *PTHLH* gene is represented below by a blue arrow. The mid-late insertion site is indicated by dotted lines surrounding a non-scaled β^A-globin ectopic origin scheme. Black numbers on the X-axis are chromosome coordinates on galGal5. Two origins sites (ORI L and ORI R) are indicated and used for further RT determination. (B) RT profiles of chromosomal alleles in DT40 Wt cells. Nascent strands (NS) were quantified by real-time qPCR in four S phase fractions. The endogenous *β-globin* locus was analysed as an early-replicated control. Specific primer pairs amplifying on both alleles determine the RT profile of the region from 140kb upstream (-140kb) to 150kb downstream (+150kb) From Hassan-Zadeh *et al*, 2012. (C) RT program along a portion of chromosome 1 is shown. The profile shown was obtained with micro-array by combining intra- and inter-array replicates with the error model algorithm for four hybridization experiments. Cells were sorted into two fractions (early and late S-phase). Immunoprecipitated BrdU pulse-labelled nascent DNA from the early and late fractions was differentially labelled and cohybridized with a chicken whole-genome oligonucleotide microarray, at a density of one probe per 5.6 kb. The log2-ratio (early/late) of the abundance of each probe in early and late S-phase is shown. Typical early, mid-late and late replication domains are shown in red, green and blue, respectively. The site of insertion is indicated (Insertion Site; black triangle).

Supplementary Figure 2: Recapitulation of SNS enrichments of β^A-globin mutant origins from Valton *et al*, 2014

(A) Scheme of the β^A-globin origin coupled with *Il2R* gene, SV40 polyA sequence and USF binding sites at scale. The different elements are represented with a particular colour code, red for the pG4#1, blue for the CACC-box, light green for the CCAAT-box, brown for the TATA-box, dark green for the sequence from CCAAT to TATA-boxes. Black bars with numbers are amplicons used for SNS quantification. Blue arrows represent the 5' deletion (relative to the pG4#1) and the orange arrows the 3' deletion. (B) Relative SNS enrichments in clones with deletions of the 5' part of the β^A-globin origin (blue), the pG4#1 (red) and the 3' part of the β^A-globin origin (orange box are shown. Errors bars indicate the standard deviation for SNS enrichment in two independent clones and qPCR duplicates.

Supplementary Figure 3: Detailed protocol of the timing shift assay from Hassan-Zadeh *et al*, 2012

(A) Simplified cartoon of the heterozygous β^A-globin origin construct insertion at the PTHLH locus. The detailed composition of this construct is described in Figure 1. Amplicons used for RT shift quantification are shown with coloured bars: "With" amplicon quantifies nascent DNA from the construct, "without" amplicon, nascent DNA from wild-type allele and "both" from homozygous alleles downstream of the insertion site. (B) Overview of the allele-specific timing shift assay experimental procedure. After clone selection, 1h BrdU pulse-labelled cells were sorted into four fractions depending on their S-phase progression from early (S1) to late (S4). For each fraction, the newly synthesized DNA strands (NS) are immunoprecipitated by anti-BrdU antibody, purified and quantified using real-time qPCR with previously described primer sets. The endogenous *β-globin* locus was analysed as an early-replicated control. To determine the difference in RT between the wt and modified alleles, we calculated the -ΔL+ΔE value, corresponding to a combination of the difference in NS enrichments quantified with the "With" and the "Without" primer sets in the S1 fraction (ΔE) and in the S3 and S4 fraction (ΔL). Both amplicon is a control for the quantification that should give an average of with and without enrichments.

Supplementary Figure 4: RT analysis of the ΔCCAAT β^A-globin origin and ΔCCAAT+TATA β^A-globin origin

(A-B) Schemes of the β^A-globin ΔCCAAT origin (A) and the β^A-globin ΔCCAAT +TATA origin (B) coupled with the *IL2R* gene, SV40 PolyA sequence and USF binding sites. The corresponding sequences are given below each scheme with pG4s underlined and *cis*-regulatory elements delineated by coloured boxes. Black line with number above represents the amplicon used for the RT Shift assay. RT profiles of chromosomal alleles for two independent clones are shown. Nascent strands (NS) were quantified by real-time qPCR in four S phase fractions. Specific primer pairs determine the RT profile for the modified allele (With), the wt allele (Without) and both alleles (Both) as described in figure S1. The endogenous *β-globin* locus was analysed as an early-replicated control. Difference -ΔL+ΔE values calculated at the target site following transgene integration are indicated. Error bars correspond to the standard deviation for qPCR duplicates.

Supplementary Figure 5: RT analysis of the ΔTATA β^A-globin origin mutant

Scheme of the β^A-globin ΔTATA origin coupled with the *IL2R* gene, SV40 PolyA sequence and USF binding sites. The corresponding sequence is given below with pG4s underlined and *cis*-regulatory elements delineated by coloured boxes. Black line with number above represents the amplicon used for the RT Shift assay. RT profiles of chromosomal alleles obtained on five independent clones are shown below. Nascent strands (NS) were quantified by real-time qPCR in four S phase fractions. Specific primer pairs determine the RT profile for the modified allele (With), the wt allele (Without) and both alleles (Both) as described in figure S1. The endogenous *β-globin* locus was analysed as an early-replicated control. Difference -ΔL+ΔE values calculated at the target site following transgene integration are indicated. Error bars correspond to the standard deviation for qPCR duplicates.

Supplementary Figure 6: RT analysis of the ΔCACC and ΔCCAAT to TATA β^A-globin origin mutants.

(A-B) Scheme of the β^A-globin ΔCACC origin (A) and the β^A-globin ΔCCAAT to TATA origin coupled with the *IL2R* gene, SV40 PolyA sequence and USF binding sites. The corresponding sequence is given below each scheme with pG4s underlined and *cis*-regulatory elements delineated by coloured boxes. Black line with number above represents the amplicon used for the RT Shift assay. RT profiles of chromosomal alleles for two independent clones are shown for each condition. Nascent strands (NS) were quantified by real-time qPCR in four S phase fractions. Specific primer pairs determine the RT profile for the modified allele (With), the wt allele (Without) and both alleles (Both) as described in

figure S1. The endogenous *β-globin* locus was analysed as an early-replicated control. Difference -ΔL+ΔE values calculated at the target site following transgene integration are indicated. Error bars correspond to the standard deviation for qPCR duplicates.

Supplementary Figure 7: RT analysis of the ΔCCAAT + TATA mpG4 β^A-globin origin mutant.

Scheme of β^A-globin ΔCCAAT +TATA mpG4#3 origin coupled with the *IL2R* gene, SV40 PolyA sequence and USF binding sites. The corresponding sequence is given with pG4s underlined and *cis*-regulatory elements delineated by coloured boxes. Stars are G to A mutations made to affect pG4#3 formation. Black line with number above represents the amplicon used for the RT Shift assay. RT profiles of chromosomal alleles obtained on two independent clones are show below. Nascent strands (NS) were quantified by real-time qPCR in four S phase fractions. Specific primer pairs determine the RT profile for the modified allele (With), the wt allele (Without) and both alleles (Both) as described in figure S1. The endogenous *β-globin* locus was analysed as an early-replicated control. Difference -ΔL+ΔE values calculated at the target site following transgene integration are indicated. Error bars correspond to the standard deviation for qPCR duplicates.

Supplementary Figure 8: RT analysis of the β^A-globin minimal origin

Scheme of the β^A-globin minimal origin coupled with the *IL2R* gene, SV40 PolyA sequence and USF binding sites. The complete sequence of the minimal origin is given with the pG4s underlined and *cis*-regulatory elements delineated by coloured boxes. Black line with number above represents the amplicon used for the RT Shift assay. RT profiles of chromosomal alleles obtained on five independent clones are shown below. Nascent strands (NS) were quantified by real-time qPCR in four S phase fractions. Specific primer pairs determine the RT profile for the modified allele (With), the wt allele (Without) and both alleles (Both) as described in figure S1. The endogenous *β-globin* locus was analysed as an early-replicated control. Difference -ΔL+ΔE values calculated at the target site following transgene integration are indicated. Error bars correspond to the standard deviation for qPCR duplicates.

Supplementary Figure 9: RT analysis of the ΔpG4#1 inverted pG4#2 β^A-globin origin.

Scheme of the β^A-globin ΔpG4#1 inverted pG4#2 origin coupled with the *IL2R* gene, SV40 PolyA sequence and USF binding sites. The corresponding sequence is given with pG4s underlined and *cis*-regulatory elements delineated by coloured boxes. Black line with number above represents the amplicon used for the RT Shift assay. RT profiles of chromosomal alleles obtained on four independent clones are shown below. Nascent strands (NS) were quantified by real-time qPCR in four S phase fractions. Specific primer pairs determine the RT profile for the modified allele (With), the wt allele (Without) and both alleles (Both) as described in figure S1. The endogenous *β-globin* locus was analysed as an early-replicated control. Difference -ΔL+ΔE values calculated at the target site following transgene integration are indicated. Error bars correspond to the standard deviation for qPCR duplicates.

Supplementary Figure 10: RT analysis of the Med14 minimal origin

Scheme of Med14 minimal origin coupled with the *IL2R* gene, SV40 PolyA sequence and USF binding sites. The corresponding sequence is given with pG4s underlined and in red boxes. Black line with number above represents the amplicon used for the RT Shift assay. RT profiles of chromosomal alleles obtained on two independent clones are shown below. Nascent strands (NS) were quantified by real-time qPCR in four S phase fractions. Specific primer pairs determine the RT profile for the modified allele (With), the wt allele (Without) and both alleles (Both) as described in figure S1. The endogenous *β-globin* locus was analysed as an early-replicated control. Difference -ΔL+ΔE values calculated at the target site following transgene integration are indicated. Error bars correspond to the standard deviation for qPCR duplicates.

Supplementary Figure 11: The β^A-globin complete and minimal origins do not drive transcription

Relative quantification by real-time qPCR of *Il2R* mRNA expression levels (RT+) or background levels (RT-) was performed in clones containing either the complete β^A-globin or the β^A-globin minimal origins. Numbers 1 and 2 represent the two independent clones used in this analysis. NA stands for non-amplified samples and NS for non-specific amplification. (A) Relative quantification values. The *Med14* gene was used as a reference gene actively transcribed and *Med14* mRNA levels were artificially set to 100. The obtained relative value for the *IL2R* gene is indicated. (B) Crossing point values for the β^A-globin complete origin and the β^A-globin minimal origin are indicated. The dilution used for the quantification is indicated in parentheses.

Supplementary Figure 12: Flow cytometry analysis of cells containing the β^A-globin minimal origin after synchronisation by elutriation and L-mimosine incubation.

Flow cytometry analysis for different fractions of elutriated cells were made after DNA labelling with propidium iodide (PI). The DNA content distributions of cells are represented for each condition. G1 phase cells are shown in green and G2/M cells in red. R3, R4 and R5 gates were settled based on the asynchronous cell profile to define the proportion of cells in each fraction. Proportions of cells analysed for each condition as well as statistics of the distribution of the detected PI values for each gate are given in the table below each graph. Black titles on the Y-axis give the name of the cell population analysed, the number of the fraction collected during the elutriation is indicated in brackets.

Supplementary Figure 13: Flow cytometry analysis of cells containing the β^A-globin inverted pG4#1 minimal origin after synchronisation by elutriation following by either L-mimosine incubation or 3 hrs of release

Flow cytometry analysis for different fractions of elutriated of cells were made after DNA labelling with propidium iodide (PI). The DNA content distributions of cells are represented for each condition. G1 phase cells are shown in green and G2/M cells in red. R3, R4 and R5 gates were settled based on the asynchronous cell profile to define the proportion of cells in each fraction. Proportions of cells analysed for each condition as well as statistics of the distribution of the detected PI values for each gate are given in the table below each graph. Black titles on the Y-axis give the name of the cell population analysed, the number of the fraction collected during the elutriation is indicated in brackets.

Supplementary Figure 14: Validation of MNase digestion patterns obtained to analysed nucleosome positioning

Chromatin was extracted from clonal cell lines containing the β^A-globin minimal origin (top gels) and the β^A-globin minimal origin inverted pG4#1 (bottom gels) and partially digested with exponentially increasing concentrations of micrococcal nuclease (MNAse; 2.5, 10, 40 and 160 U/mL). The four digested DNA samples obtained for each clonal cell line were subjected to electrophoresis in a 1% w/v agarose gel and stained with SYBR safe. The DNA size marker was a commercial 1 kb plus ladder. Digestion patterns of cells synchronised at the G1/S transition are shown on the left and digestion patterns of cells synchronised in G2 are shown on the right.

Supplementary Figure 15: Nucleosome positioning over the β^A-globin minimal origin analysed by real-time qPCR

Scheme of the β^A-globin minimal origin coupled to the *IL2R* gene (partly shown) and associated with two upstream USF binding sites. Upstream, relative enrichments of DNA found inside mono-nucleosomes purified after MNase digestion of the corresponding clone are shown. G1/S transition synchronised cells are represented in blue and the G2 synchonised cells are represented in red. Numbers (from +1 to +5) represent the positioned nucleosomes relative to the pG4#1.

Supplementary Figure 16: Genome-wide Gs tracks distribution around pG4s in an origin or non-origin context

(A) G triplets distribution around pG4s that are not associated to origin (green line) or associated with Ori (Yellow line). Median is given for each data sets and the significance of the difference is assessed using p-value. (B) G triplets distribution around pG4 associated with origin activity in reads per kilo base per million mapped reads (rpkm). Black bars are the total pG4 distributed over the genome, green bars are non-origin associated pG4s and yellow bars are origin associated pG4s. (C) G triplets distribution around pG4s over replication origin associated or not with CGI.

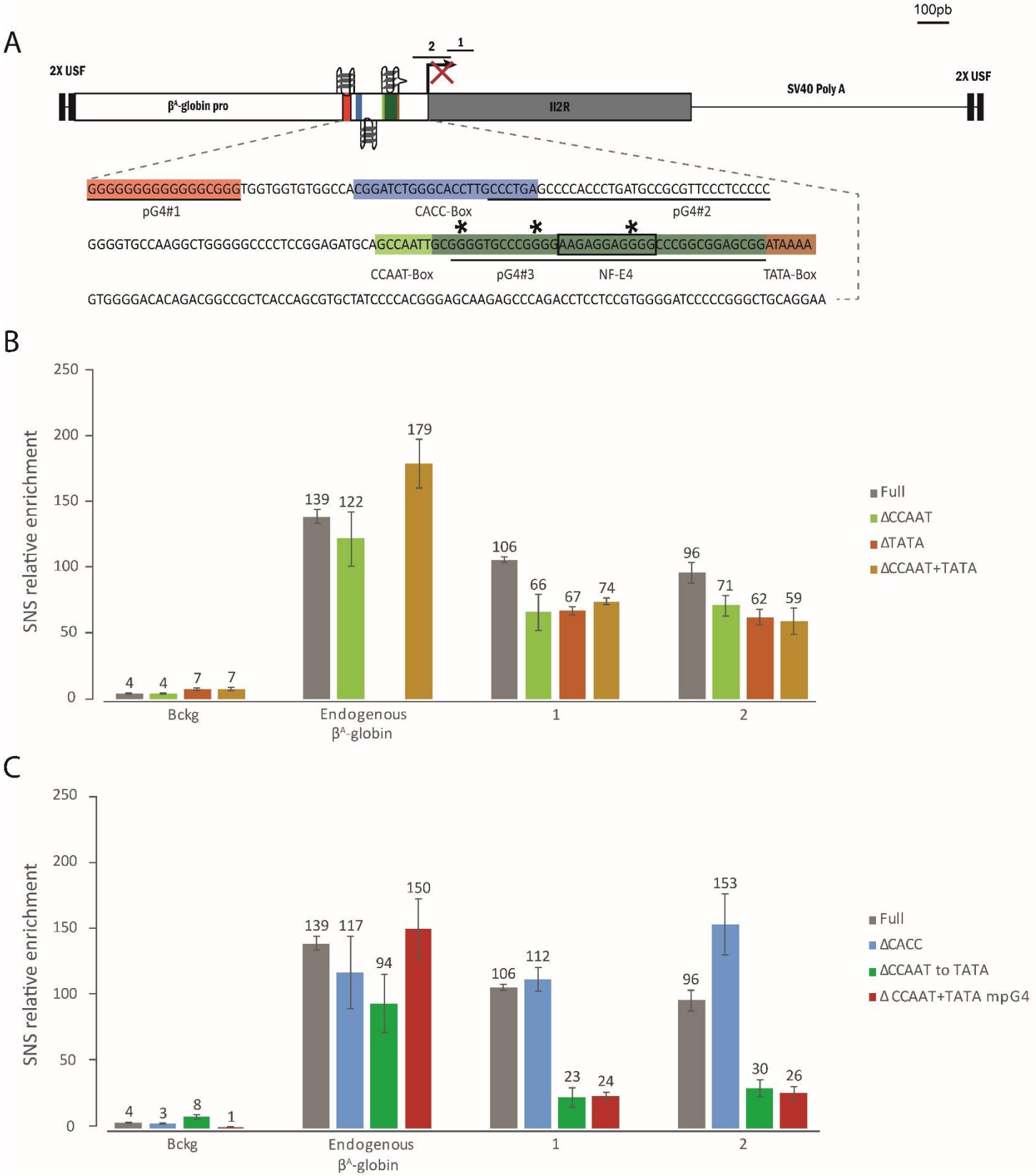


Figure 1: Identification of essential *cis*-elements inside the β^A-globin origin

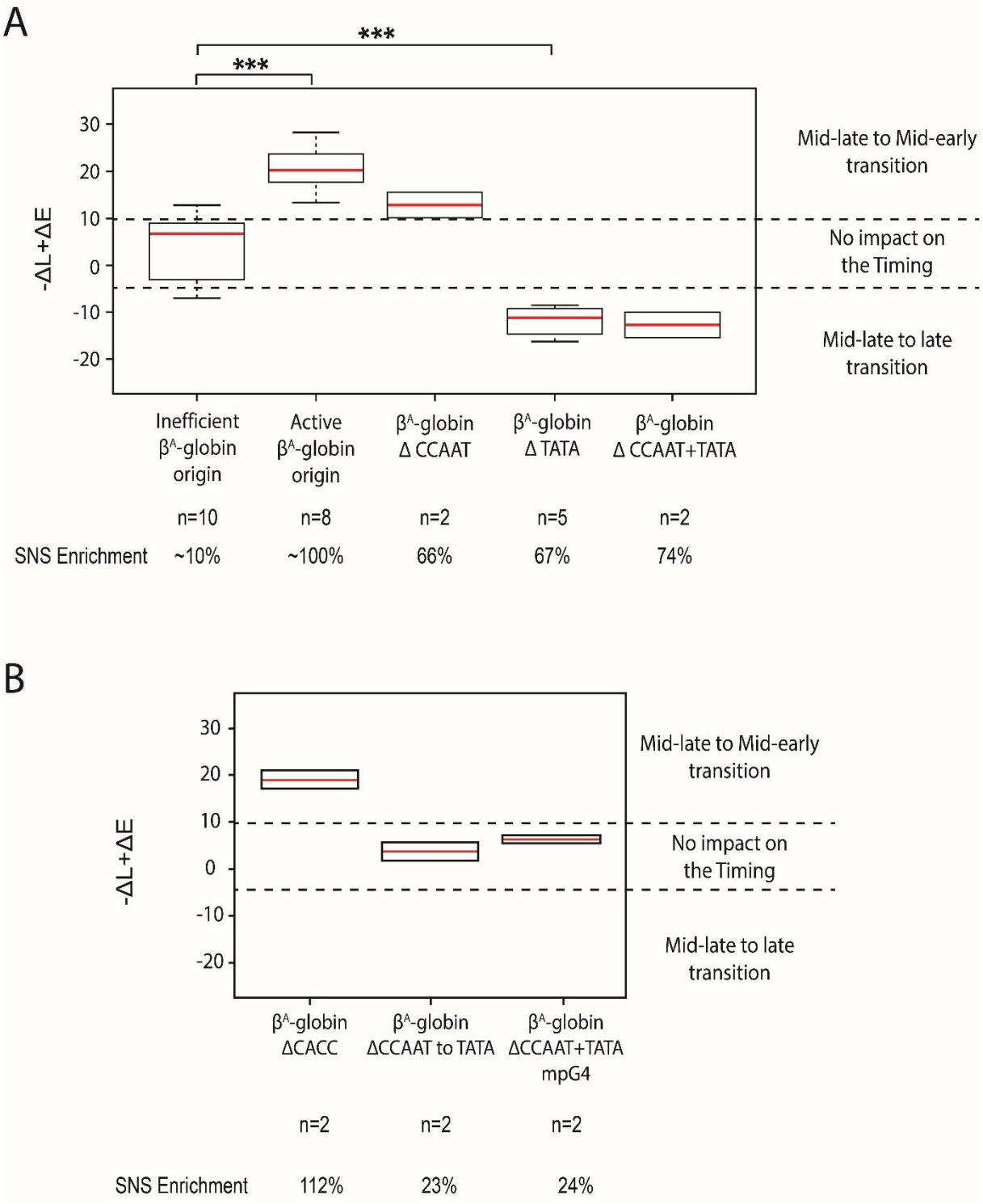


Figure 2: RT shift assay reveals an unexpected role of the TATA-box on timing

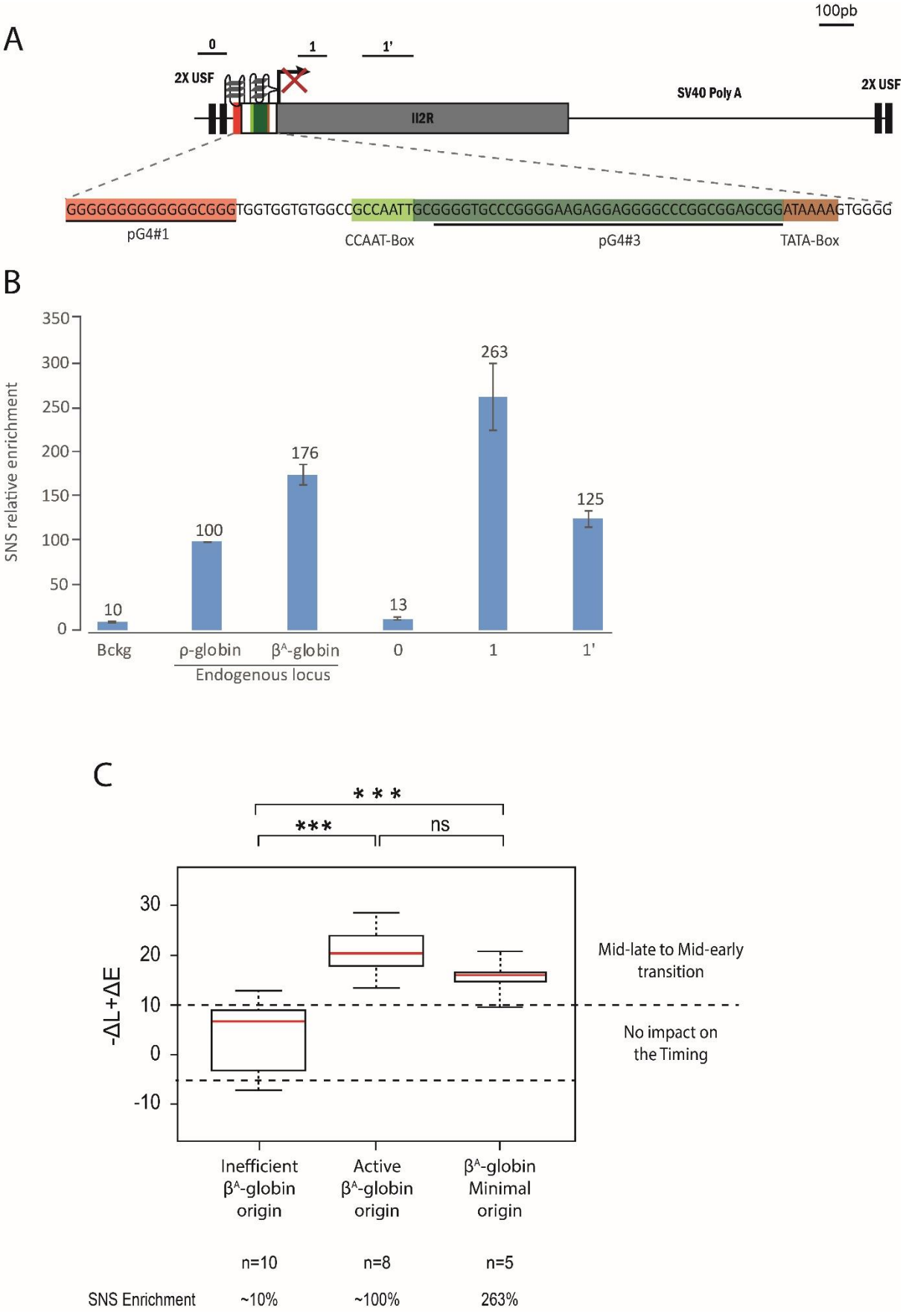


Figure 3: A 90 bp β^A-globin minimal origin drives efficient DNA replication initiation

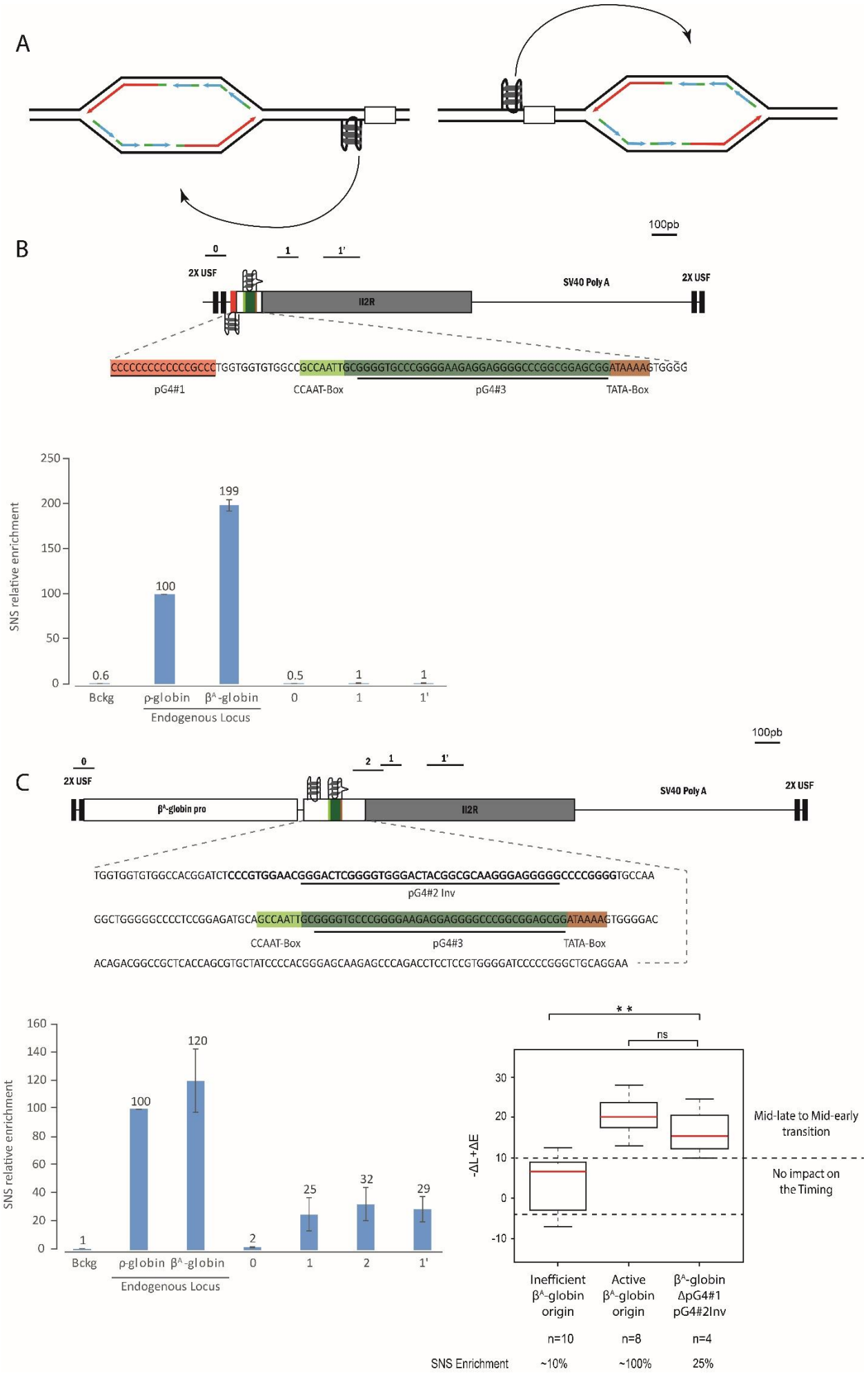


Figure 4: pG4 orientation impacts on the activity of the minimal origin

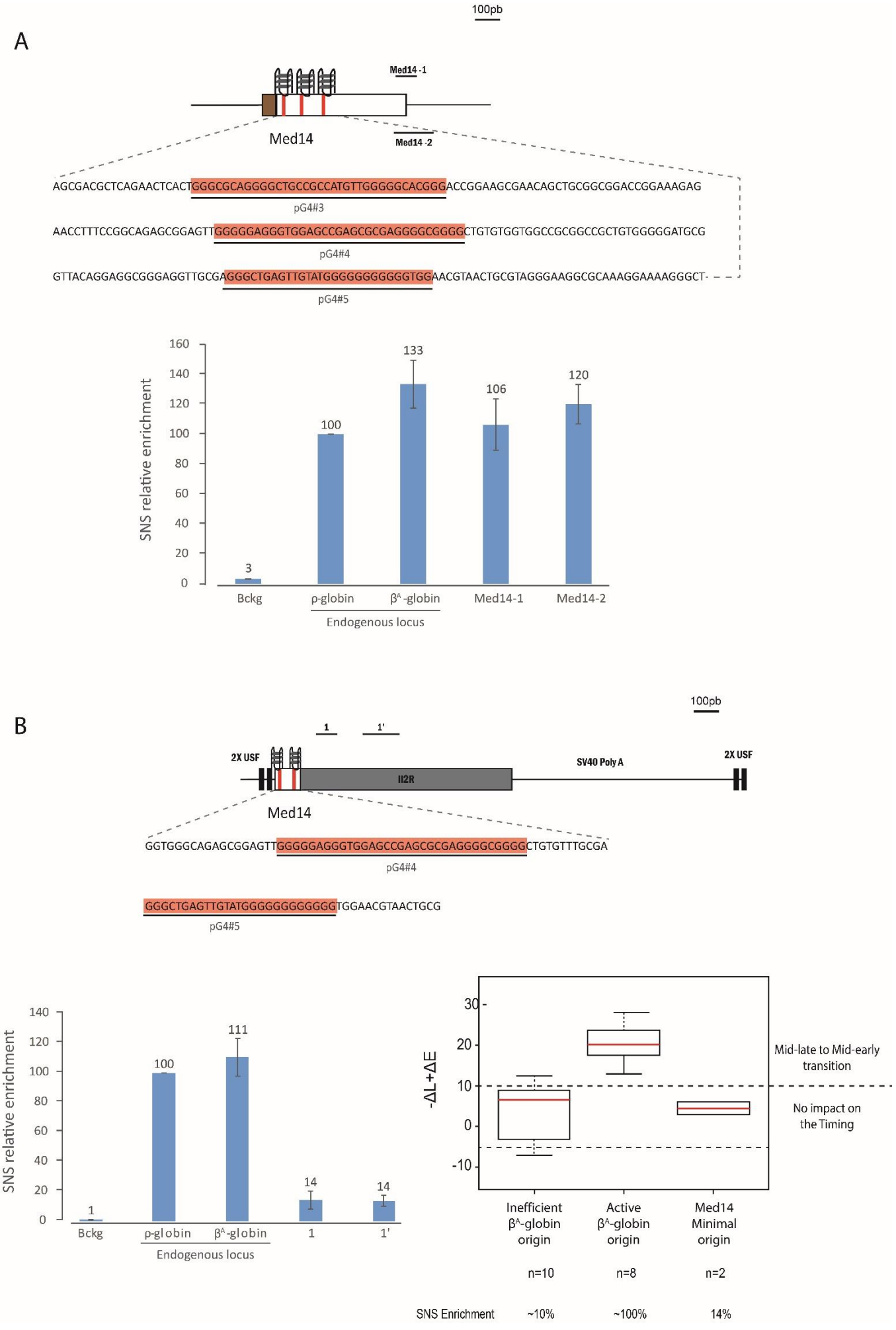


Figure 5: Two pG4s found inside the Med 14 origin are not sufficient to form an active

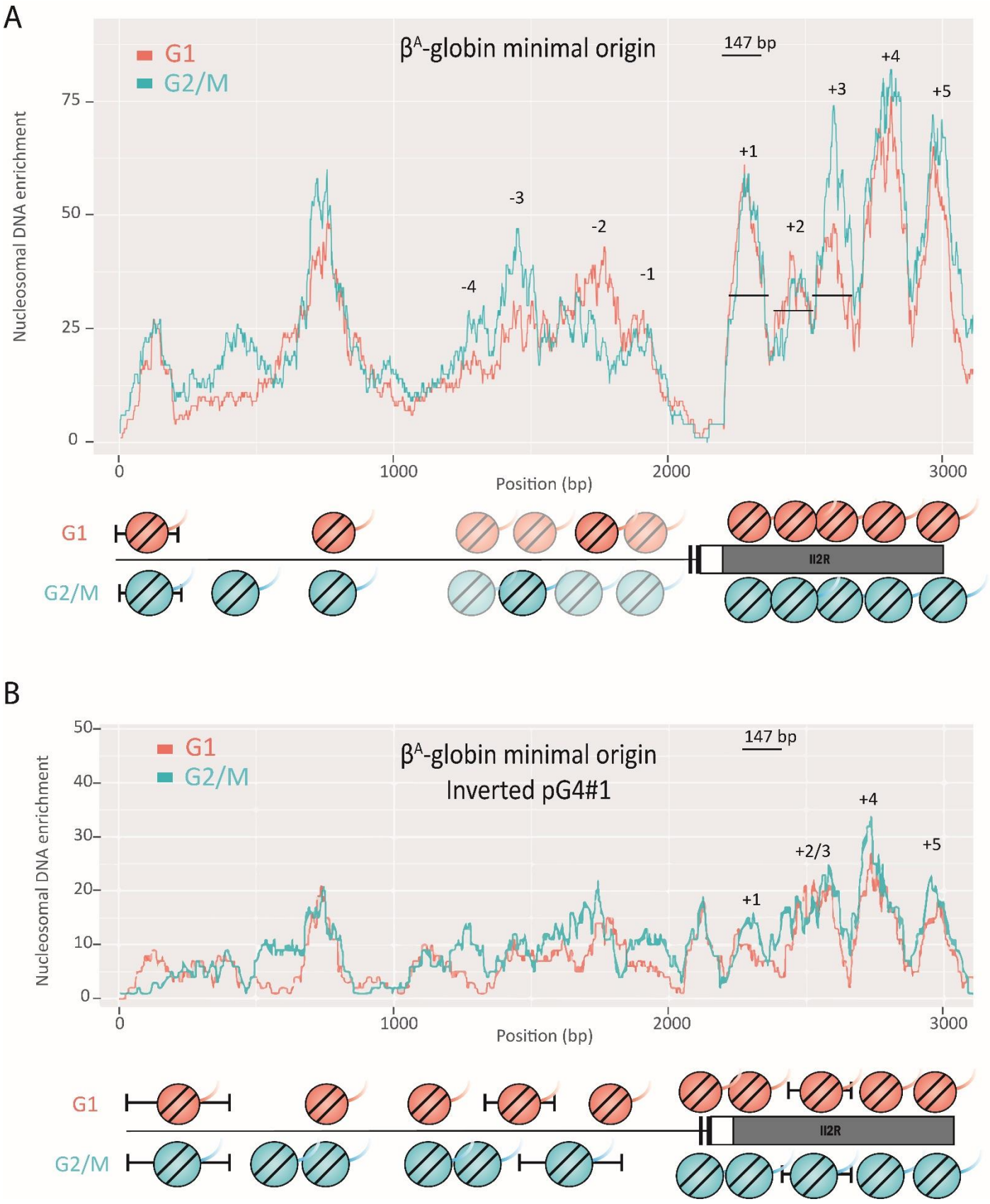


Figure 6: The β^A-globin minimal origin induces a precise nucleosome organisation

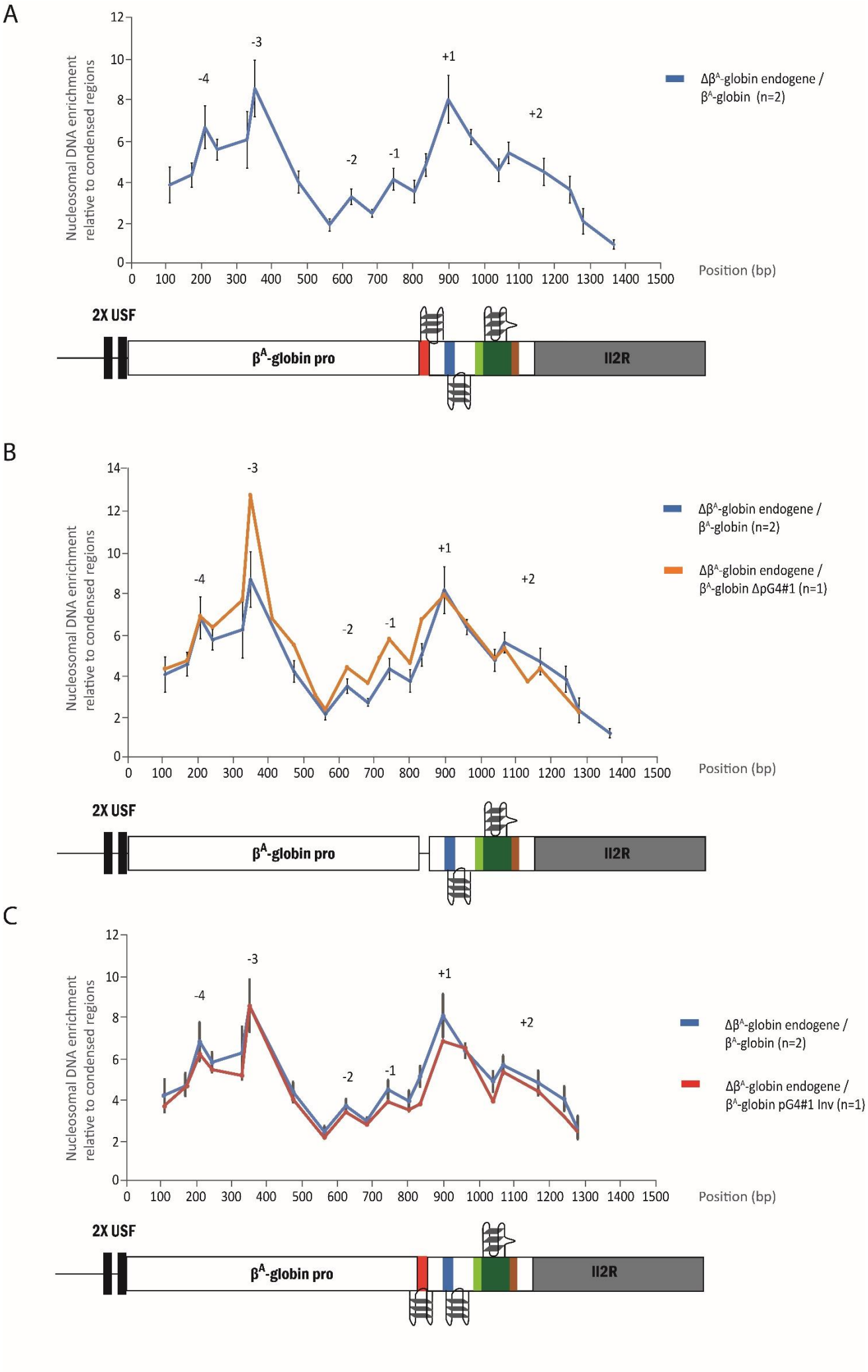


Figure 7: Nucleosome positioning over the β^A-globin origin also reveals a specific pattern

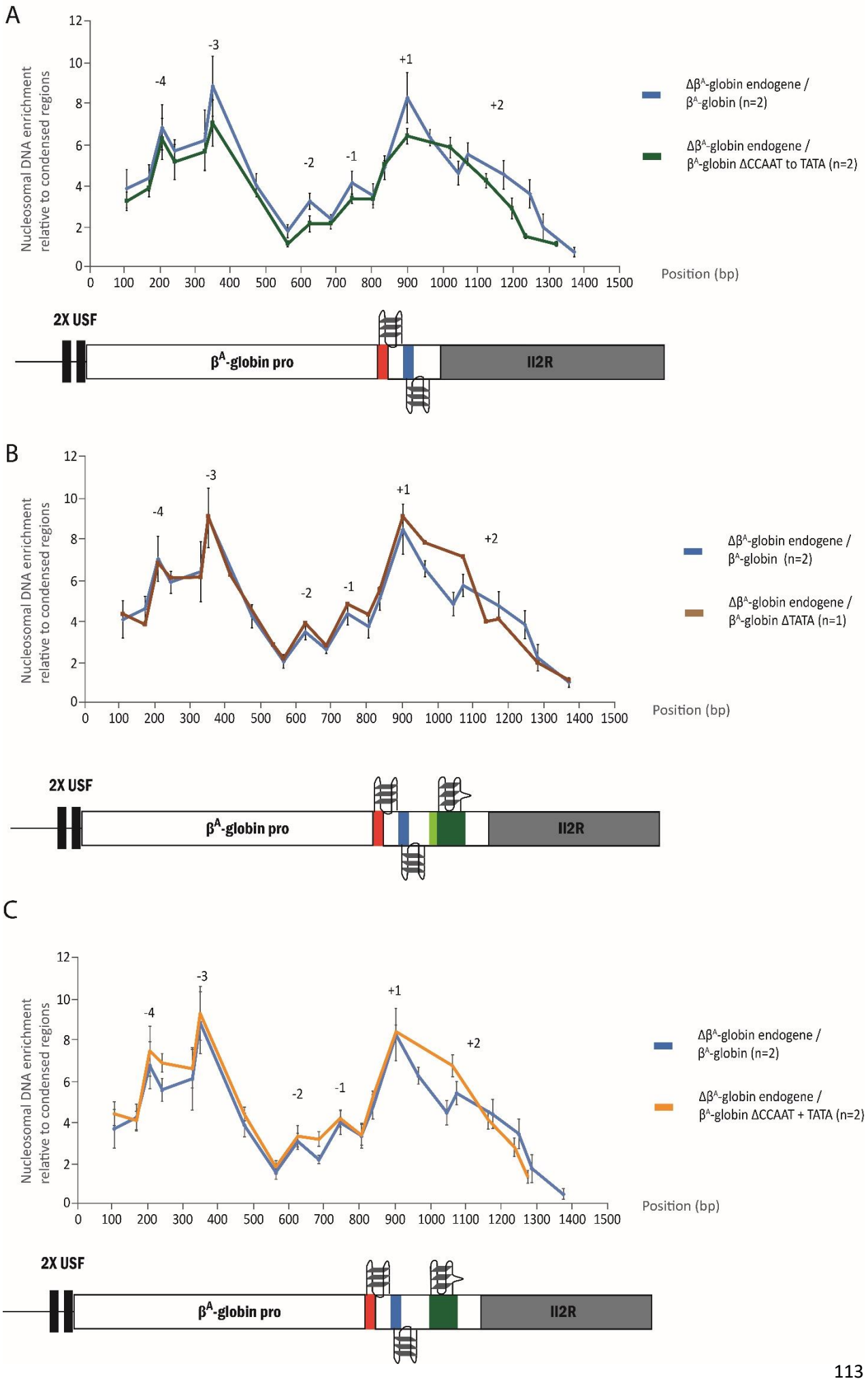
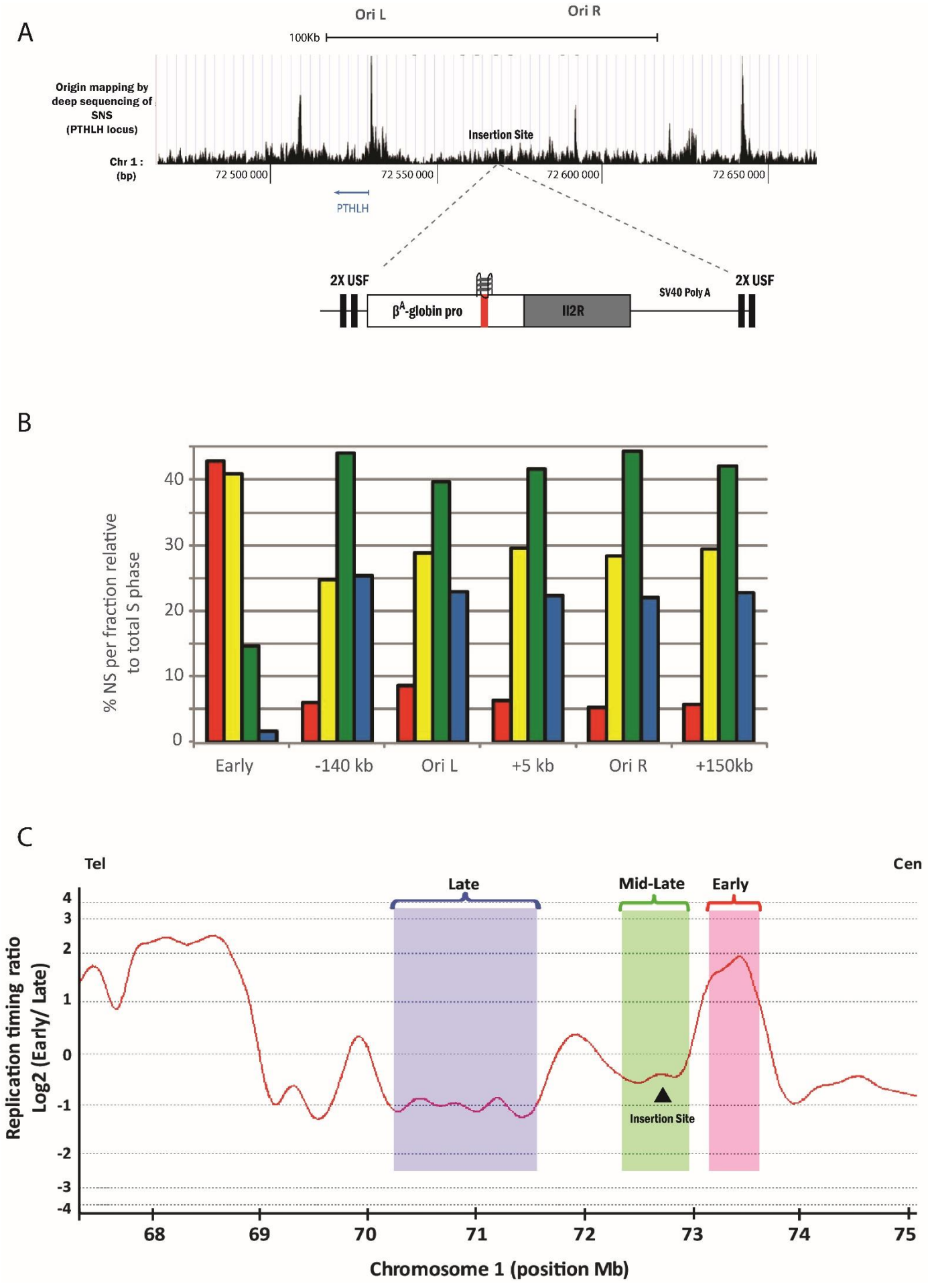
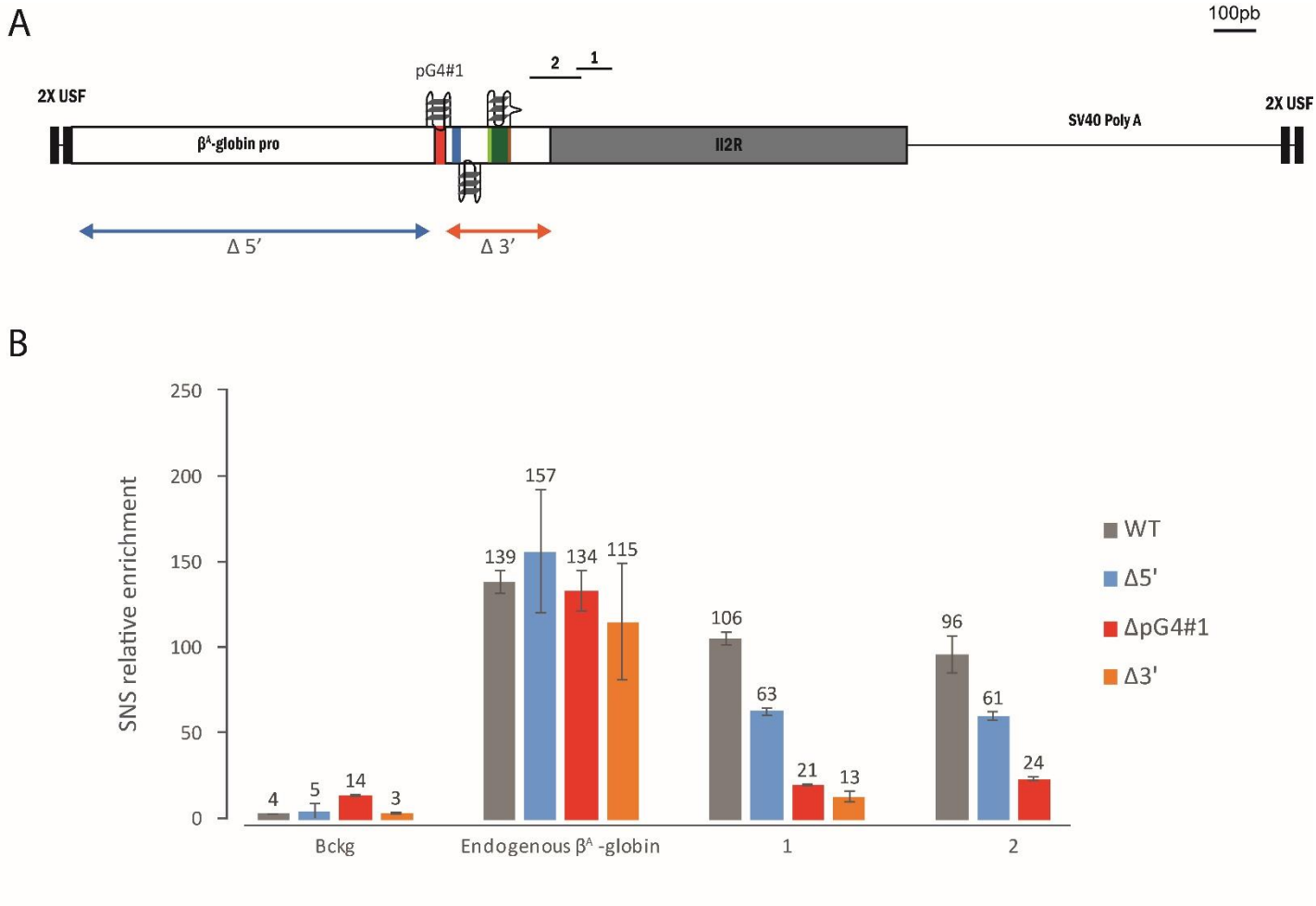


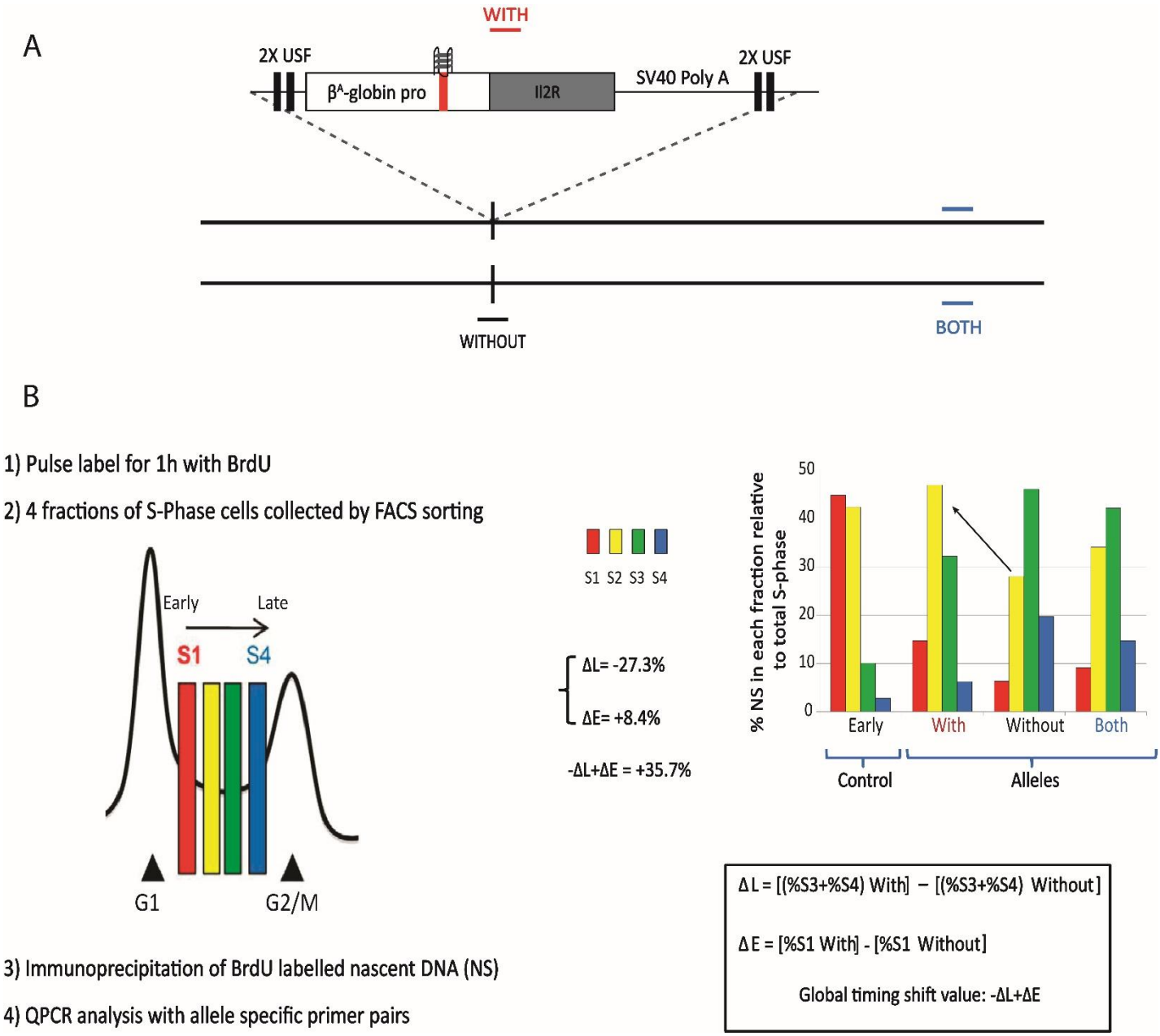
Figure 8: The TATA-box is involved in nucleosome positioning



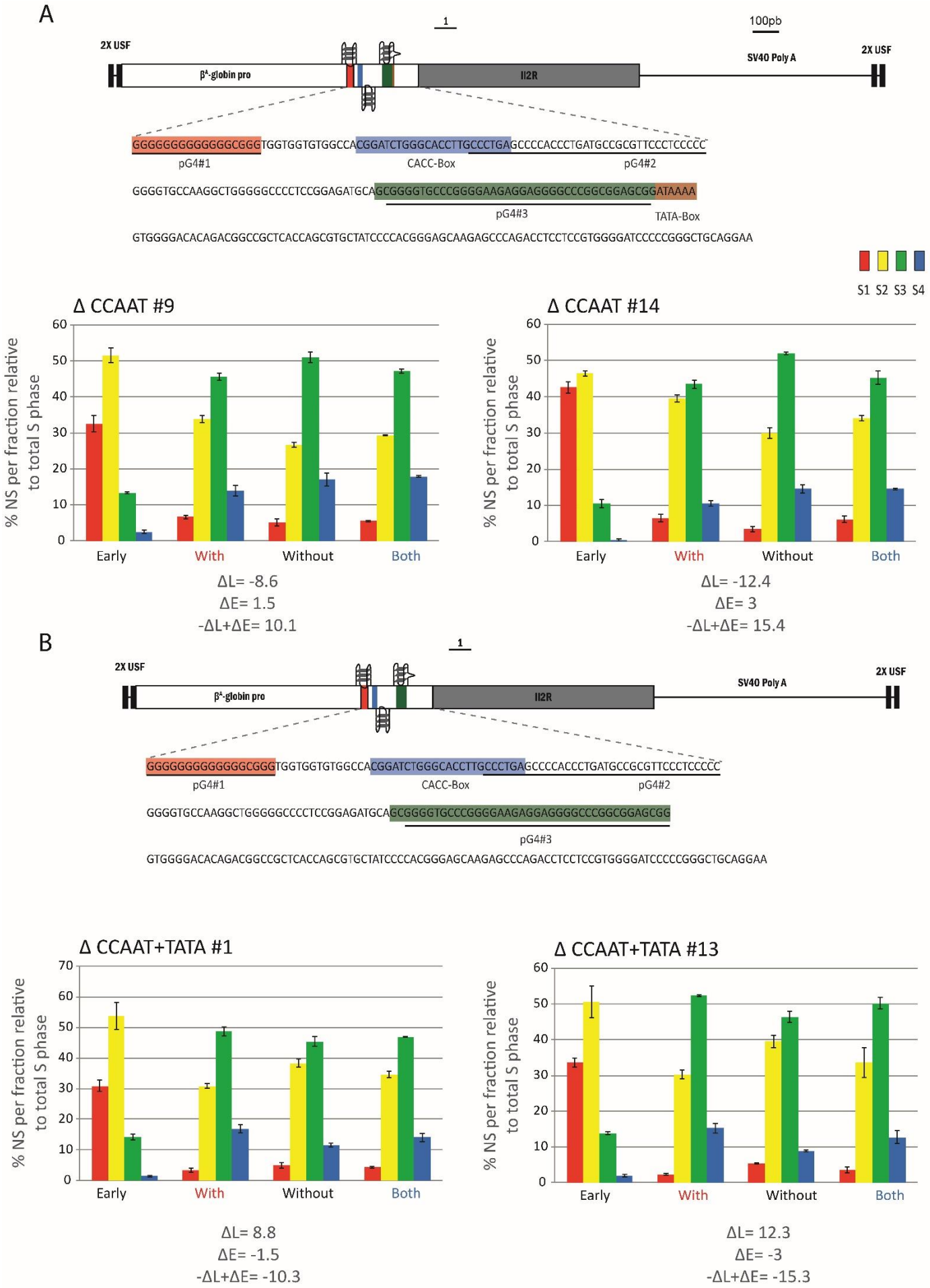
Supplementary Figure 1: Properties of the insertion site used in this study



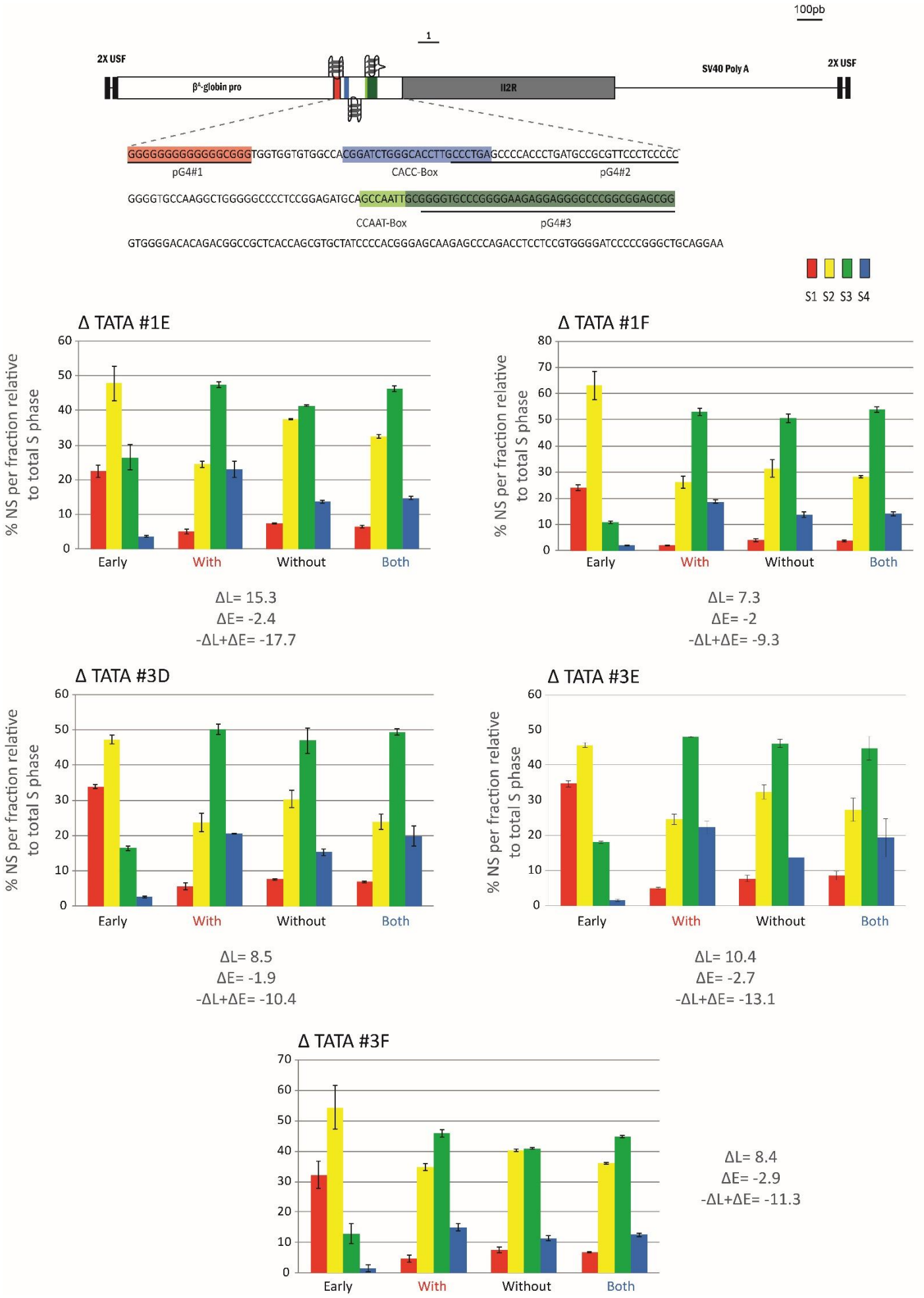
Supplementary Figure 2: Recapitulation of SNS enrichments of β^A-globin mutant origins from Valton *et al*, 2014



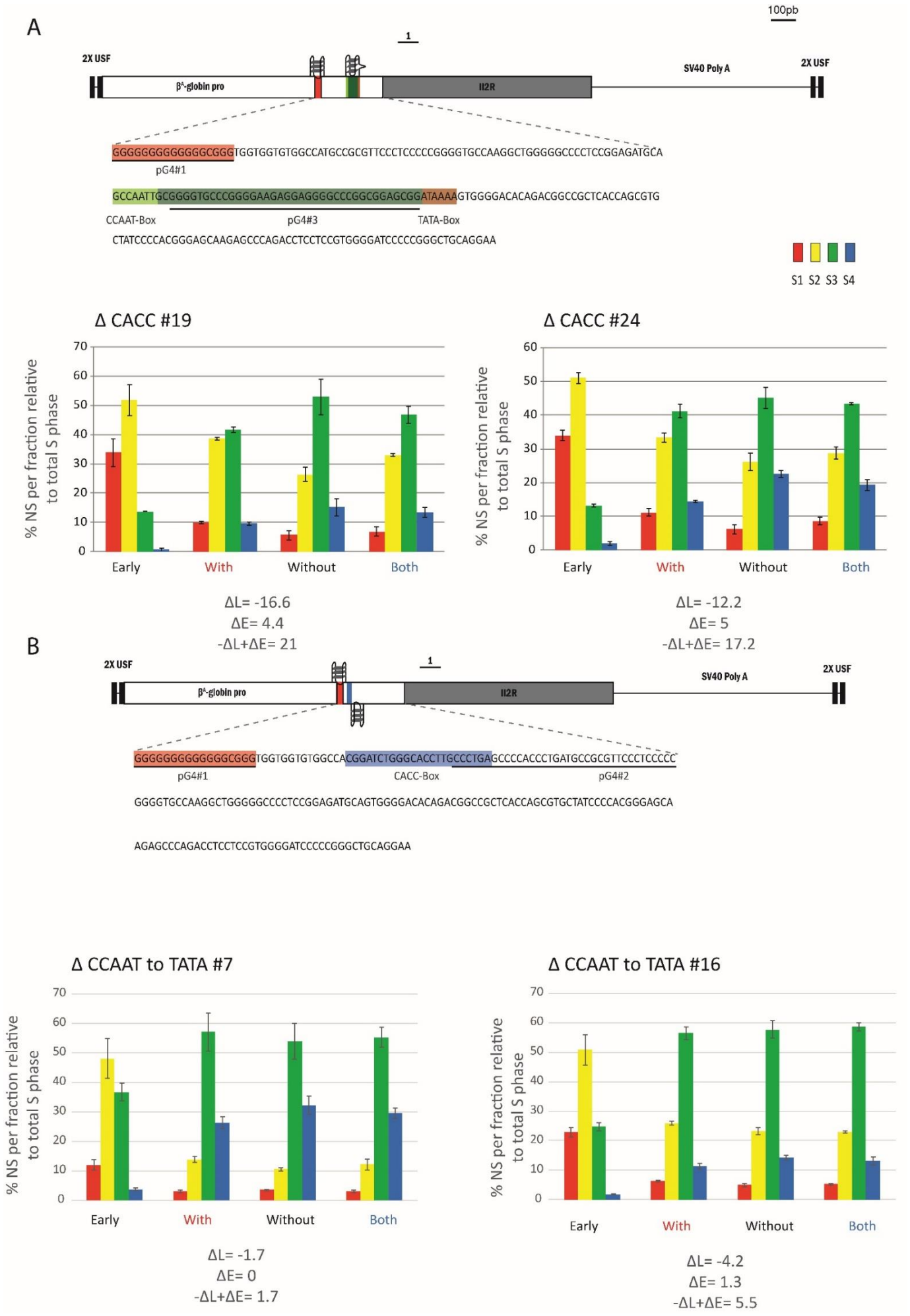
Supplementary Figure 3: Detailed protocol of the timing shift assay from Hassan-Zadeh *et al*, 2012



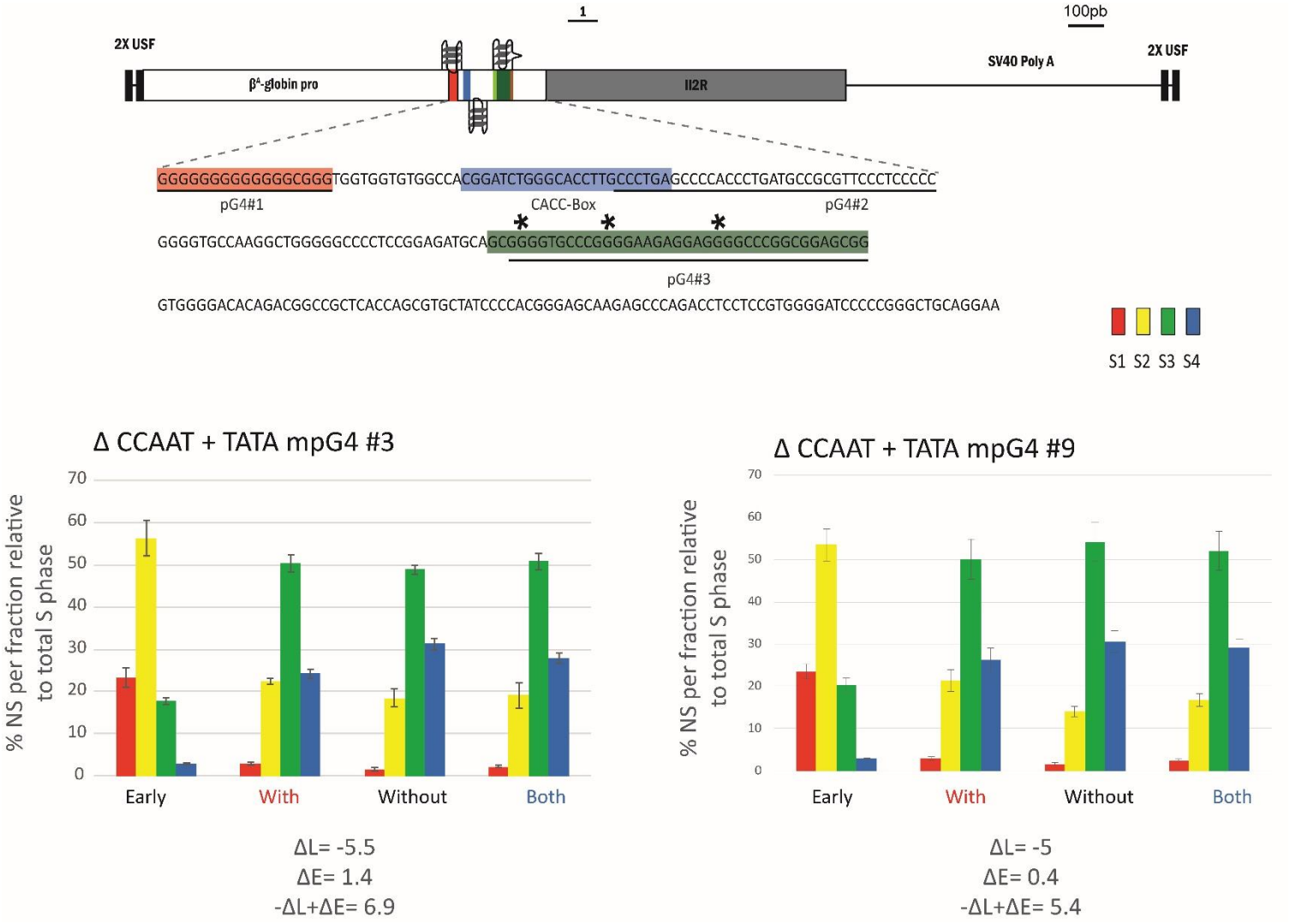
Supplementary Figure 4: RT analysis of the ΔCCAAT β^A-globin origin and ΔCCAAT+TATA β^A-globin origin



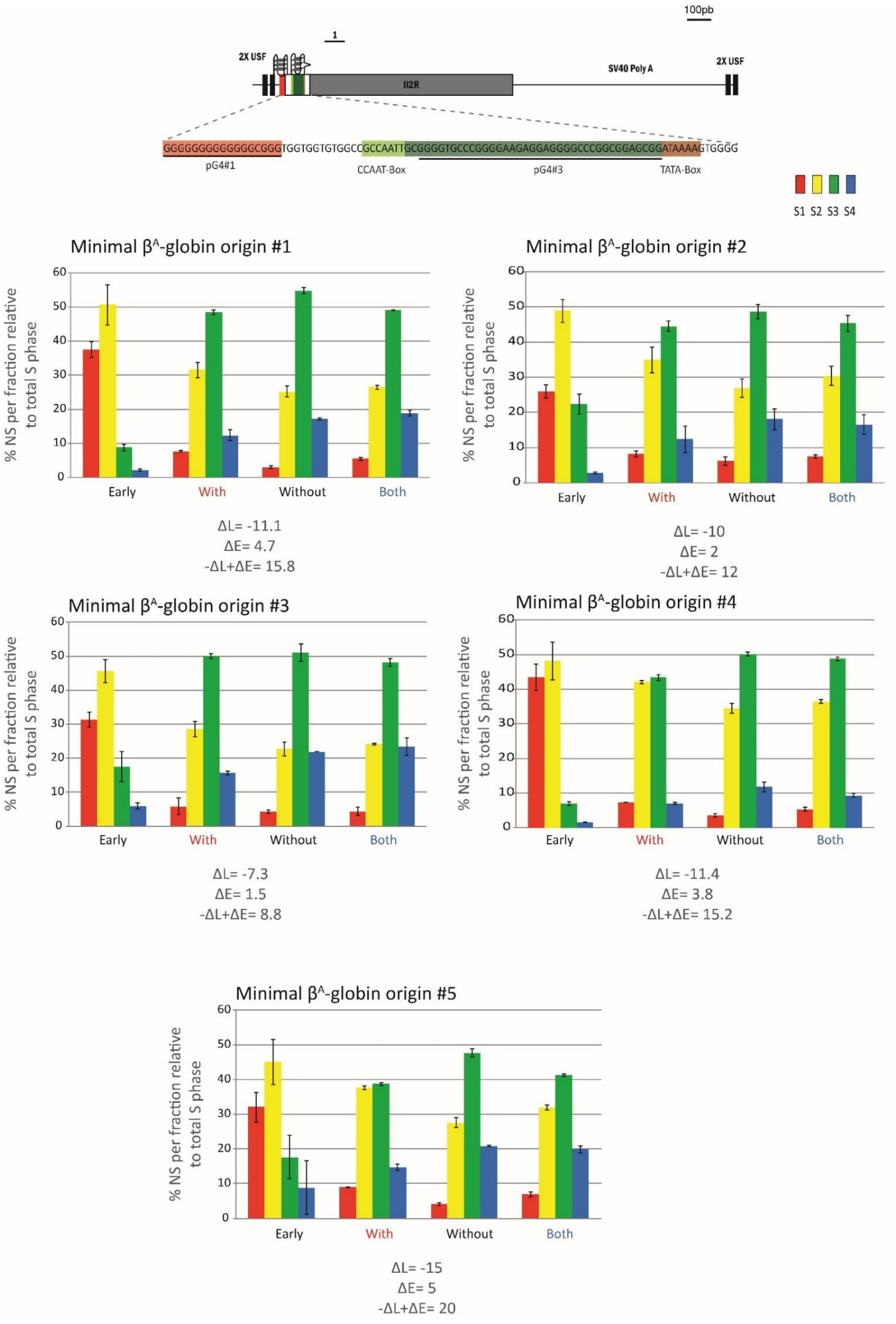
Supplementary Figure 5: RT analysis of the ΔTATA β^A-globin origin mutant.



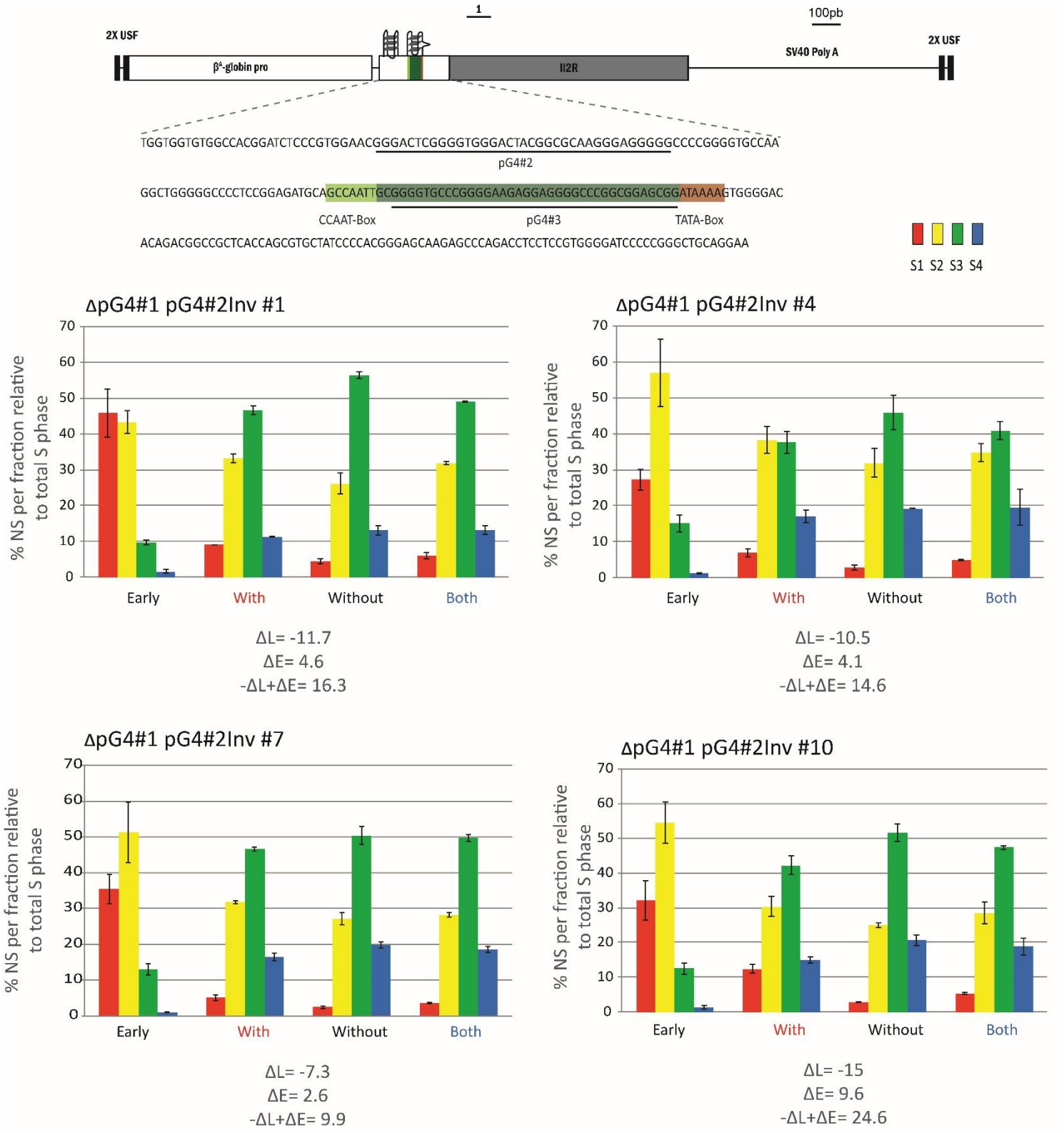
Supplementary Figure 6: RT analysis of the ΔCACC and ΔCCAAT to TATA β^A-globin origin mutants.



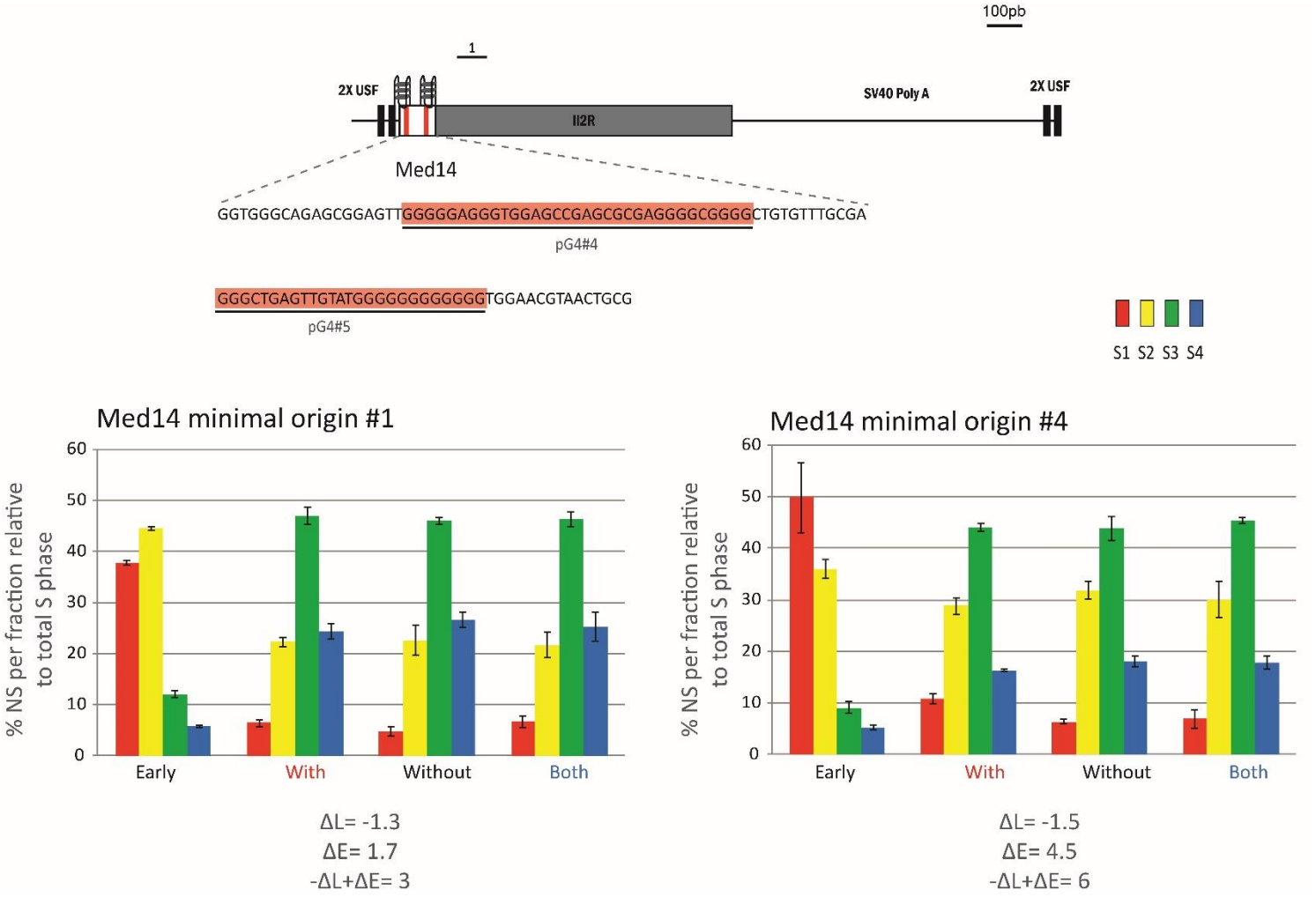
Supplementary Figure 7: RT analysis of the ΔCCAAT + TATA mpG4 β^A-globin origin mutant.



Supplementary Figure 8: RT analysis of the β^A-globin minimal origin



Supplementary Figure 9: RT analysis of the ΔpG4#1 inverted pG4#2 β^A-globin origin. β^A-



Supplementary Figure 10: RT analysis of the Med14 minimal origin.

A

Relative quantification values		*Med14* gene	*Il2R* gene
RT+	β^A-globin (1)	100	0.04
	β^A-globin (2)	100	0.03
RT+	β^A- Minimal globin (1)	100	0.04
	β^A-Minimal globin (2)	100	0.01

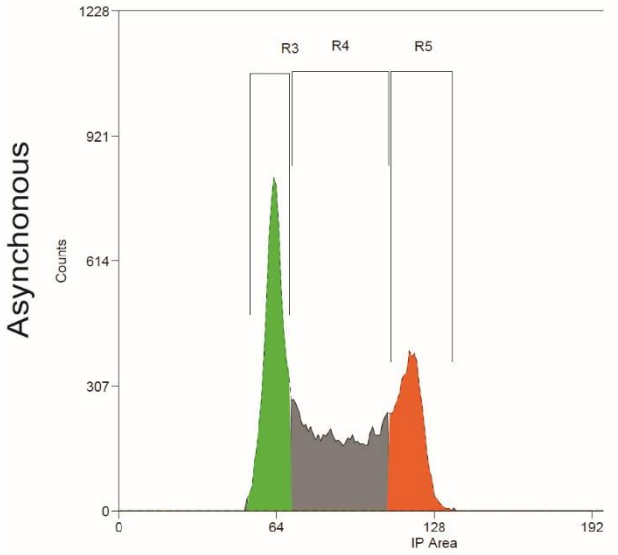
B

Crossing point values		*Med14* gene
RT+	β^A-globin (1) (1/100)	28.93
	β^A-globin (2) (1/100)	28.03
RT-	β^A-globin (1)	NA
	β^A-globin (2)	NA
RT+	β^A Minimal globin (1) (1/100)	28.68
	β^A Minimal globin (2) (1/100)	27.68
RT-	β^A Minimal globin (1)	NA
	β^A Minimal globin (2)	NA

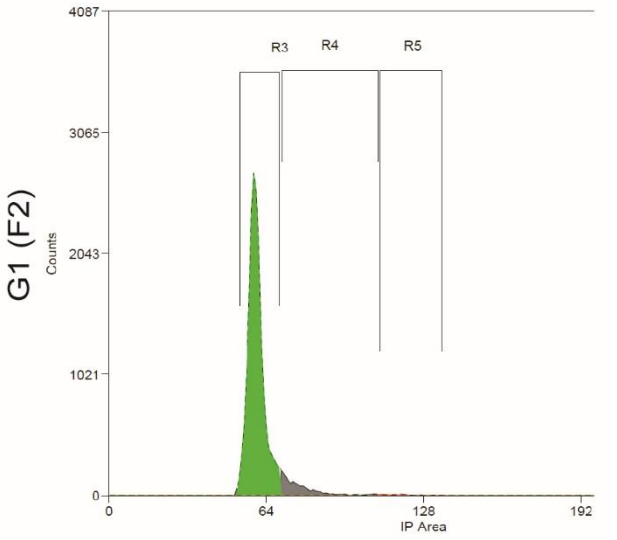
Crossing point values		*Il2R* gene
RT+	β^A-globin (1)	35.41
	β^A-globin (2)	35.14
RT-	β^A-globin (1)	NA
	β^A-globin (2)	NA
RT+	β^A Minimal globin (1)	35.27
	β^A Minimal globin (2)	36.46
RT-	β^A Minimal globin (1)	NS
	β^A Minimal globin (2)	NA

Supplementary Figure 11: The β^A-globin complete and minimal origins do not drive transcription

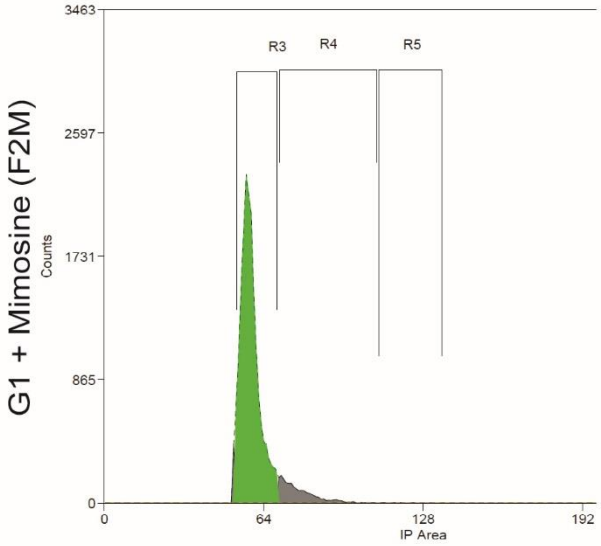
Synchronised β^A-globin minimal origin



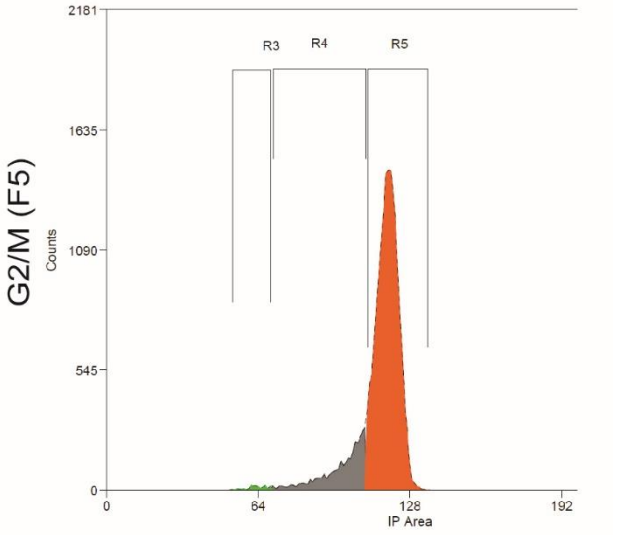
Region	Count	% Hist	Median	Mean	CV
Total	20000	100.00	83.00	86.83	26.34
R3	7151	35.75	63.00	62.82	5.62
R4	7789	38.95	88.00	88.80	13.68
R5	5024	25.12	118.00	118.07	4.03



Region	Count	% Hist	Median	Mean	CV
Total	20000	100.00	60.00	61.83	12.92
R3	18009	90.05	60.00	59.84	5.29
R4	1801	9.01	75.00	77.97	10.95
R5	128	0.64	118.00	118.34	5.27



Region	Count	% Hist	Median	Mean	CV
Total	20000	100.00	58.00	60.55	12.62
R3	17523	87.61	58.00	58.61	6.13
R4	2033	10.17	76.00	77.99	9.37
R5	31	0.16	117.00	119.03	4.99

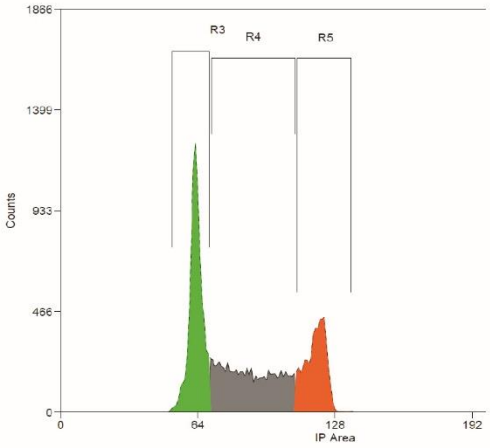


Region	Count	% Hist	Median	Mean	CV
Total	20000	100.00	118.00	114.84	9.33
R3	222	1.11	63.00	62.67	6.77
R4	3292	16.46	102.00	98.92	9.53
R5	16479	82.39	119.00	118.73	3.59

Supplementary Figure 12: Flow cytometry profiles of cells containing the β^A-globin minimal origin after synchronisation by elutriation and L-mimosine incubation.

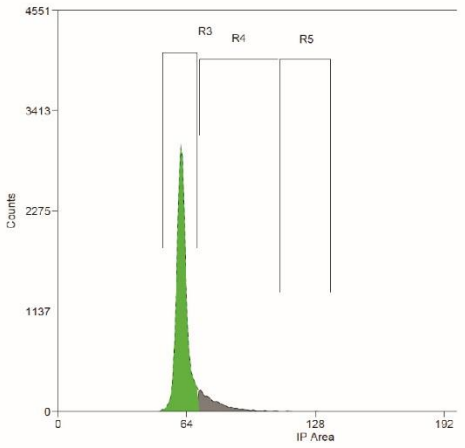
Synchronised β^A-globin minimal origin inverted pG4#1

Asynchonous



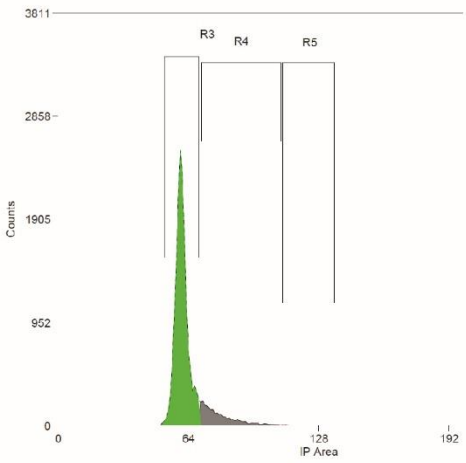
Region	Count	% Hist	Median	Mean	CV
Total	20000	100.00	80.00	86.08	26.74
R3	7704	38.52	63.00	63.02	4.78
R4	7393	36.96	88.00	88.43	13.36
R5	4896	24.48	119.00	118.82	3.76

G1 (F2)



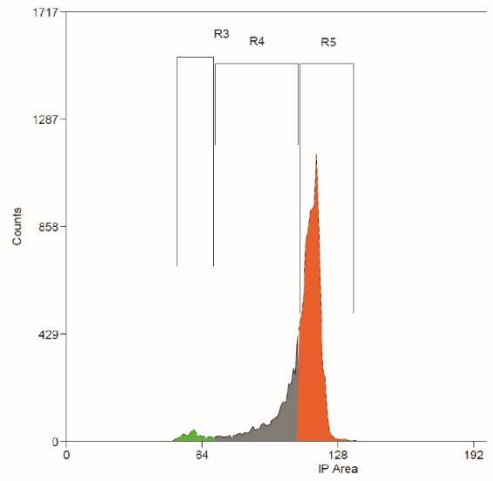
Region	Count	% Hist	Median	Mean	CV
Total	19189	100.00	62.00	63.75	10.35
R3	16785	87.47	62.00	61.71	4.31
R4	2379	12.40	76.00	77.65	9.31
R5	20	0.10	116.00	119.30	5.35

G1 + Mimosine (F2M)



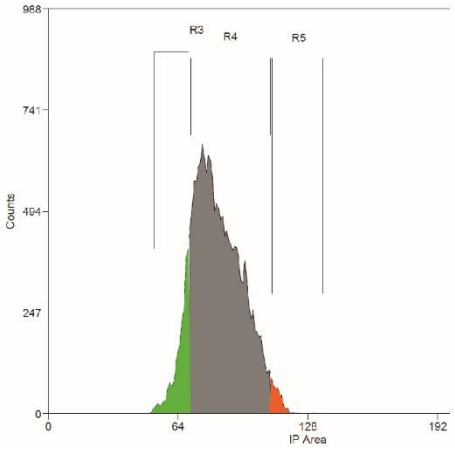
Region	Count	% Hist	Median	Mean	CV
Total	18386	100.00	61.00	63.26	12.47
R3	15790	85.88	60.00	60.64	5.06
R4	2554	13.89	77.00	79.13	10.62
R5	23	0.13	112.00	114.09	4.12

G2/M (F5)



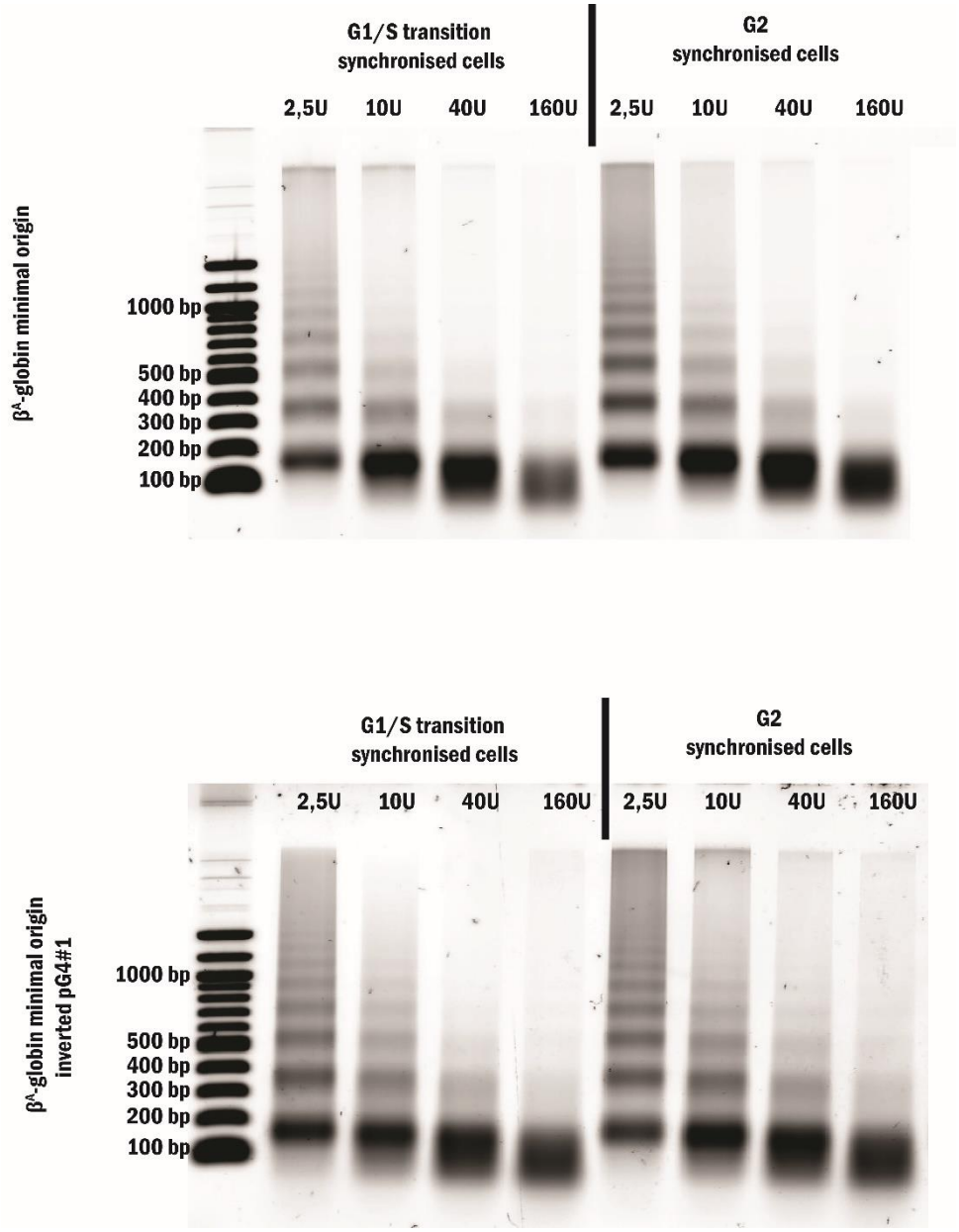
Region	Count	% Hist	Median	Mean	CV
Total	13150	100.00	114.00	109.93	11.82
R3	422	3.21	60.00	60.27	7.58
R4	3406	25.90	103.00	99.48	9.67
R5	9312	70.81	116.00	116.04	2.96

G1 + 3h (F2+3h)

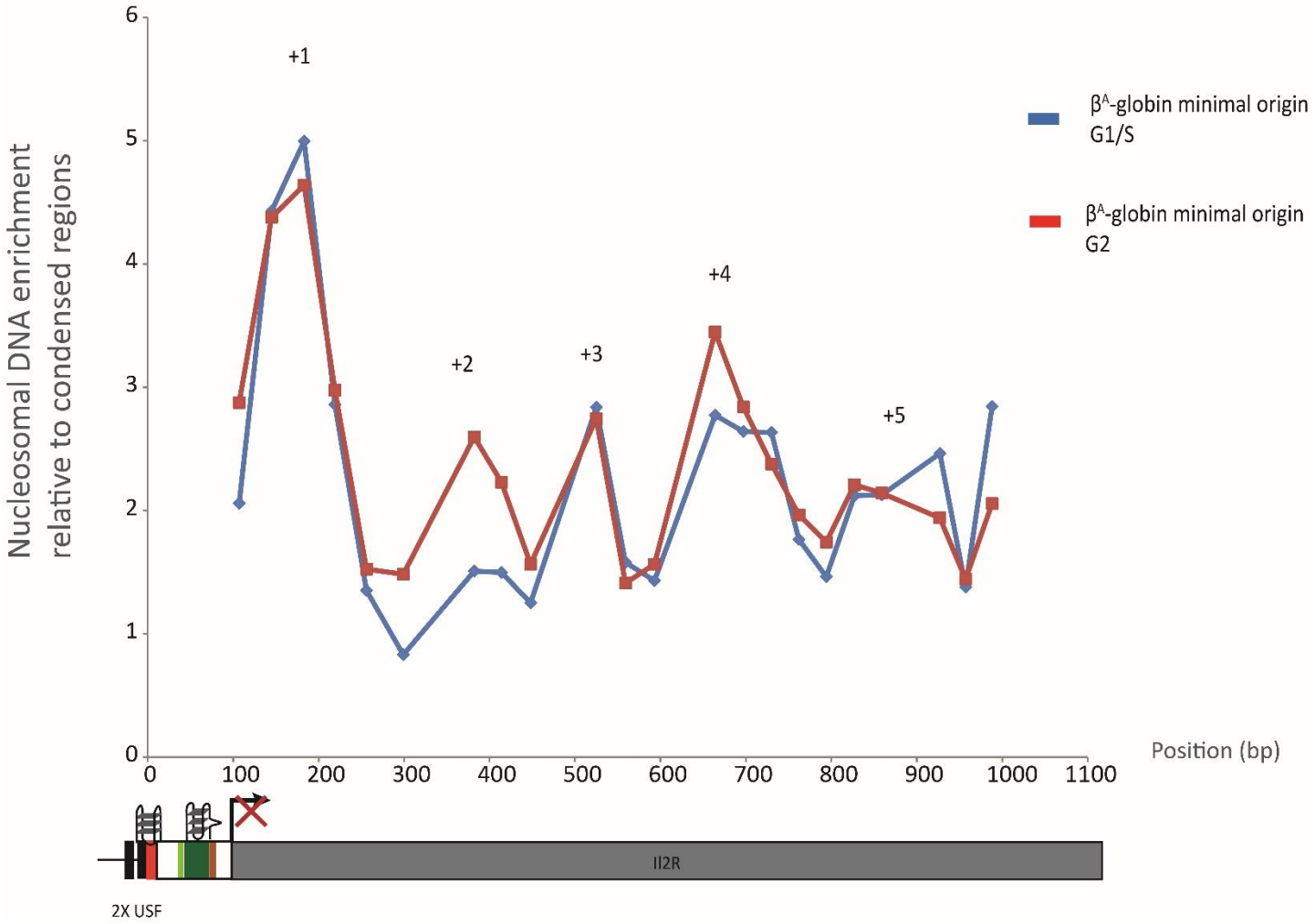


Region	Count	% Hist	Median	Mean	CV
Total	18685	100.00	82.00	83.42	14.74
R3	2161	11.57	66.00	65.16	5.96
R4	16093	86.13	84.00	85.10	11.90
R5	427	2.29	112.00	112.83	2.27

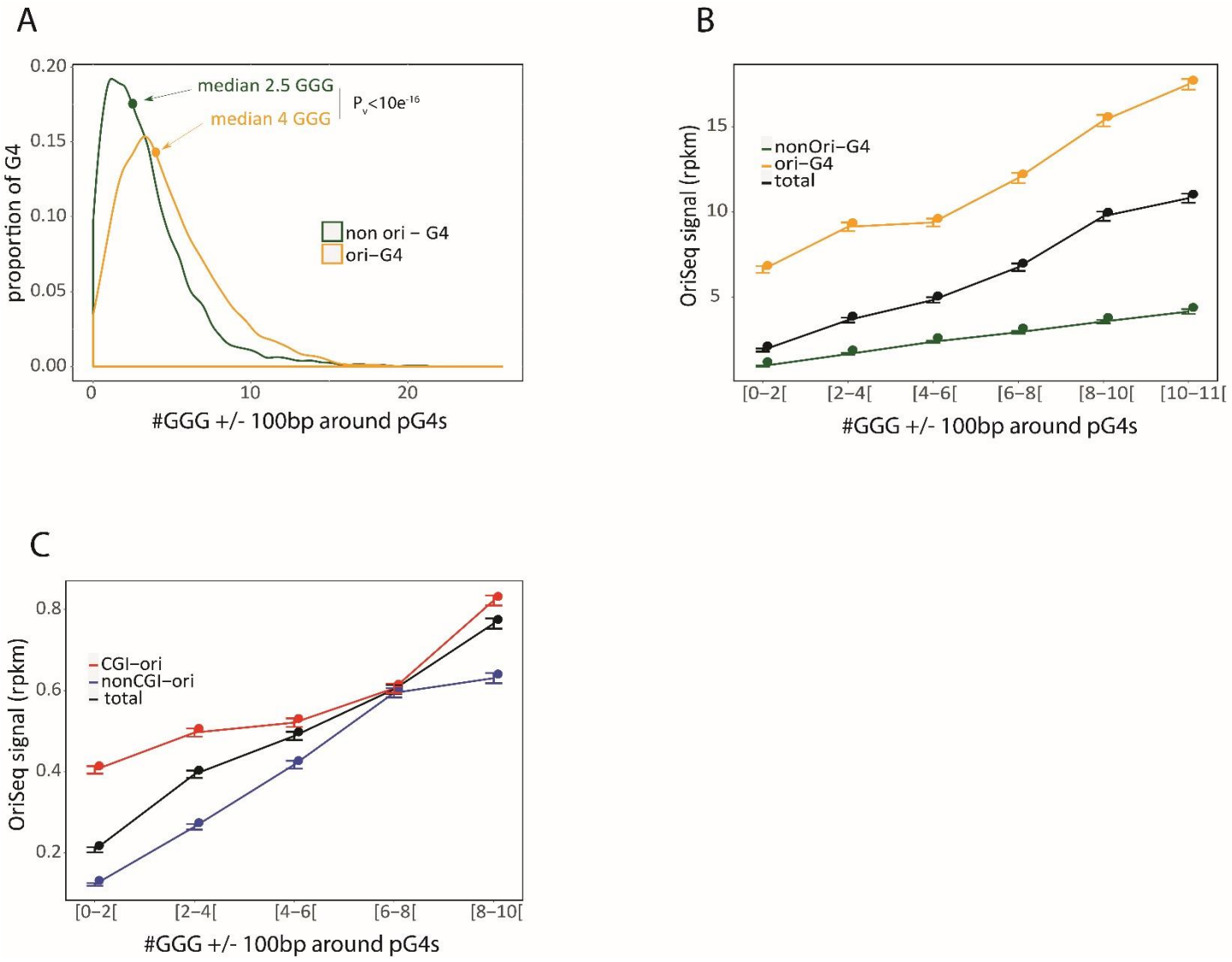
Supplementary Figure 13: Flow cytometry profiles of cells containing the minimal β^A-globin inverted pG4#1 origin synchronisation by elutriation, plus L-mimosine incubation or 3 hrs release



Supplementary Figure 14: MNase DNA digestion profiles



Supplementary Figure 15: β^A-globin minimal origin nucleosome map by qPCR



Supplementary Figure 16: Genome-wide Gs tracks distribution around pG4s in origin or not context

Part3
Discussion

TATA-box implication on replication timing

Identification of replication origins and replication timing profiles, in metazoan, genome-wide has revealed a strong correlation between transcription and replication. Efficient early replicated origin are found enriched at active promoters. Whereas late replicated and gene poor regions[56,70] are origin poor. The relationship between the replication-timing program and transcription is more complex, indeed, recent studies suggest an impact of transcription on the replication timing depending on the length of the transcribed gene. A strong promoter driving transcription of a long gene profoundly affect the RT whereas the same promoter in front of a small gene has only a mild effect. These results, however, do not address the molecular mechanism involved in this regulation. In our study, identification of the TATA-box role on the RT brings another direct evidence that transcription *cis*-regulatory elements are important components of the RT program. Indeed, deletion of the TATA-box has a strong effect on the RT control of our model origin, although, its efficiency is only midly affected. As a control, deletion of the CCAAT-box leads to a similar effect on origin efficiency but does not impact RT as the TATA-box deletion.

The TATA-box is recognised by the TBP factor, which plays a major role in the transcription machinery recruitment. It will be of major interest to determine how the TATA-box influence the replication timing. The first main interrogation is whether TBP is recruited or not to the TATA-box. The underlying question is, whether TBP is required to influence the RT or does the TATA-box alone is sufficient to induce a change? A first way to answer this question would be to ChIP-qPCR TBP to address its recruitment to the β^A-globin origin. However, the TBP detection could be limited by ChIP approach poor sensitivity for transient interactions. Moreover, no expression of IL2R reporter gene was detected when fused with the complete β^A-globin origin construct, it is then possible that TBP is not recruited to the origin. A direct way to test the impact of TBP recruitment on RT control would be to use a fusion of TBP with GAL4 and test whether an UAS sequence might replace the TATA-box at the active origin. UAS sequence is recognised by the GAL4 protein and will induce TBP recruitment at the origin level. However, TBP recruitment would probably trigger transcription and therefore would question, in case of an impact on RT, whether TBP or the process of transcription is involved in this regulation. We could also try to directly insert the TATA-box in an ectopic replication origin to see its behaviours on RT. However, this approach does not ensure that the TATA-box would be in the required location to influence RT. Indeed, the observation that in both the entire and the minimal β^A –globin origin the TATA-box is located at the border of a nucleosome suggests a direct role of the TATA-box on the nucleosome positioning. This hypothesis is reinforced by the result obtained on the ΔTATA-box mutant where nucleosome organisation is strongly perturbed. It is known that the primary sequence can play a critical role on nucleosome positioning, we can also hypothesised that the TATA-box sequence itself

plays a role in RT control. Finally, results obtained in the laboratory show that replication origins, when changing the replication timing, also affect its surrounding, changing replication timing of the origins found nearby (Brossas et al, in revision). It would be of great interest to make 4C experiments to compare the interaction of the WT and the ΔTATA β^A-globin origin with the surrounding region to see whether they develop different patterns of interaction.

G-rich sequence potentially forming a G4

Extensive analysis of the β^A-globin origin essential *cis*-regulatory module identifies a G-rich sequence potentially forming a G4, which is predicted to structure into a G4 containing a bulge[84]. It will be important to test its formation and stability *in vitro* and to define mutants strongly affecting its formation without strongly changing its G-richness. Indeed, *in vivo* studies are quite fastidious and the identification of such mutant would allow to generate sequences of similar G-richness that are able, or not to structure into a G4 and see how the origin behaves in these conditions. Similarly to our previous study on pG4#1, such analysis would help us to define whether or not a second G4 is necessary to form a strong origin.

Recently, a precise positioning of SNS peaks identified a 40bp window that shows only limited SNPs diversity, suggesting a selection pressure over this sequence. Search for motif enrichment inside this 40bp window revealed several motifs such as CCC(GGG), GAG(CTC) and CG(GC) found in human, mouse and chicken[69]. However, the role of these motifs has not being addressed in our study, but we cannot exclude an effect of these motifs on the origin activity. Indeed, in our case the site of initiation is found to be around 100-200 bp downstream of our defined origin.

From MNase experiments made on the complete β^A-globin origin, we concluded that a nucleosome is well positioned over the G-rich sequence pG4#3. However, it has to be kept in mind that we did not address the nature of the protecting protein and that it could be a non-histone protein[109]. If the replication firing or/and licensing require the pG4#3 structure, then, how is it formed when covered by a nucleosome ? This is why it would be interesting to perform the same experiment as those made with the minimal β^A-globin origin, with cells synchronised in G1 and G2/M and also to make ChIP with antibodies against histones after MNase digestion (will be discussed later).

Founding these two potentials G4 as essential *cis*-elements also asks the question of how they might cooperate to induce the replication initiation. It would be interesting to verify, *in vitro*, the propensity of each G4 to structure themselves in the presence of the second G4 and their resulting stability.

The distance between pG4#1 and #3 does not seem to be a major determinant of origin activity as the minimal origin is active and the distance between the two pG4 is only 22bp (G4#1 tail to pG4#3 head)

whereas it is 110bp in the complete β^A-globin origin. How can we compare this result with the observation that in *S.cerevisiae* efficient ARS tend to contain two ACS in opposite orientations, one of high-affinity and an additional less specific site? These G4 distances can be put in parallel of the in vitro/in vivo study in budding yeast, using differently spaced ARS that was still able to support DNA replication with a distance between the ACSs starting from 25bp to 400bp[65].

The β^A-globin minimal origin as a paradigm for metazoan origins

Deep investigation of the β^A-globin origin allows us to define an origin devoid of unessential sequences for the replication origin activity. Such minimal origin represents an important step in the identification of the minimal signal defining a replication origin. This minimal origin contains two pG4s with one being surrounded by CCAAT and TATA-boxes. However, several questions remain to be address. Firstly, are the CCAAT-box and the TATA-box essential for the minimal replication origin activity? We already showed that the CCAAT and TATA-boxes were not essential for the activity of the complete β^A-globin origin, however, in the minimal β^A-globin origin context, lacking most of the "auxiliary" sequences, we cannot exclude that the CCAAT and TATA-boxes have an important role on the replication origin activity. To test their role, we can simply delete these two boxes or only one after the other inside the minimal β^A-globin origin and test their impact on SNS enrichment. If the ΔTATA-box mutant maintained a strong SNS enrichment, it would be interesting to test its impact on RT. A RT delay similar to the one observed with the entire origin would suggest that the TATA-box inside the minimal origin plays the same role. Loss of the origin activity by the deletion of the CCAAT and TATA-boxes would deeply complicated our understanding of the role of *cis*-regulatory elements on origin function since depending on the context, the same deletion would lead to a different effect.

The minimal β^A-globin origin could be the starting point of *in silico* characterisation of replication origins and the development of prediction algorithm, however, a similar minimal origin (Med14 minimal origin) did not ensure the replication initiation. The Med14 minimal origin that contain two G-rich sequences, potentially forming a G4, as in the minimal β^A-globin origin did not allow to detect a strong SNS enrichment. Suggesting that this two G-rich sequence alone cannot sustain replication activity. The comparison of this sequence with the β^A-globin minimal origin, underlines the lack of the CCAAT and TATA-box sequences as well as a strong and canonical G4 sequence. As discussed before, we cannot be sure, as long as we have not test it, that the CCAAT and the TATA-boxes inside the minimal origin context are not essential. We could try to add the two sequences to the Med14 minimal sequences (in complement of β^A-globin minimal origin CCAAT and TATA-box deletion). However, a lack of origin activity would not necessarily means that the CCAAT and TATA-boxes are not essential, but could simply suggest that, in this conformation, they do not work. Secondly, the pG4s should also be

investigated to see if they can form G4 *in vitro* and what are their respective stabilities. It would be interesting to exchange one of the pG4 of the Med14 origin and to replace it by the pG4#1 of the β^A-globin origin. This pG4#1 exhibit a canonical sequence as well as a strong stability, it also seems to position the 5' border of the strongly positioned nucleosome +1 on the complete β^A-globin origin.

It would be of great interest to test our β^A-globin minimal origin in human cells and see if it can still support DNA replication initiation. The easiest way would be to do episome maintenance experiment. That would not involve any genome modification as well as SNS purification and quantification.

Dissecting the Pre-RC recruitment *in vivo*

The minimal β^A-globin origin could be the starting point of deep characterisation of pre-RC formation dynamic in metazoan. Independently of Pre-RC assembly, it remains unknown how does the Pre-RC is addressed to replication origins. *Cis*-regulatory elements essential for origin activity could be essential for the licensing, the firing, or both. So far only one study tried to address this question using a cell free model system, the Xenopus egg extract. When 80-mer oligonucleotides containing pG4s were pre-incubated in the extracts, pre-RC licensing could be observed but not the firing, suggesting a role of pG4s in firing rather than on origin licensing[73]. Using the β^A-globin minimal origin, we could also address these questions *in vivo* by directly mapping formation of pre-RCs. We plan to use the ChEC-seq technic standing for Chromatin endogenous cleavage, based on the fusion of a protein of interest with the active MNase core. Recruitment of the protein of interest on the genome will induce DNA digestion that will result in the release of small DNA fragments (around 25bp)[124]. MNase activation is induced by permeabilization of cell's membranes and addition of Ca^{2+}, without chromatin fixation, it allows a dynamic digestion of DNA nearby factors of interest binding sites. Such technic allows to address the chromatin localisation of a protein of interest independently of the use of any antibody as in ChIP technics, avoiding antibody bias caused by ChIP conditions. Moreover, the ChEC can detect weak and transient interactions in opposition to ChIP. ChEC-seq of ORC and MCM [13], in a β^A-globin minimal origin or in a pG4#1 inverted β^A-globin minimal origin context will inform us on the presence or not of the Pre-RC. Pre-RC presence on both origins (active and non-active) will indicate the Pre-RC licensing occurs on both origins no matter their activity and G4 orientations, but it will also indicate that the identified essential sequences are required for firing and not for licensing. In opposition, Pre-RC recruitment to only the β^A-globin minimal origin will show the role of *cis*-regulatory elements during the licensing step. ORC has been found as binding G4 structure, also its mapping will help deciphering its interaction pattern[100]. The mapping of both ORC and MCM sub-units should also inform us on how ORC delivers MCM onto the chromatin and test the hypothesis that one nucleosome sits in between ORC binding and MCM loading.

We could also go further in the dissection of the origin activation process by fusing firing factors to the MNase core. In case both origins are licensed, independently of their sequence specificity, CDC45[13] or MTBP[98] firing essential elements fusion with MNase will allow to differentiate which step of origin activation is impaired in function of origin sequence. MTBP is of great interest since it recognizes G4s and acts as a dimer.

Moreover, this approach can be used genome-wide, indeed, using DT40 synchronisation in G1 by elutriation and mimosine blockage, will allow, with MNase core fusion with ORC subunits and MCM, to address Pre-RC formation genome-wide. Even though these proteins have been already mapped[41–43], ChEC-seq should produce a more precise cartography of the overall distribution of loaded Pre-RCs. Such cartography, coupled with firing factors ChEC-seq (in asynchronous cells) will allow to identify active origins and could be the opportunity to dissect the Pre-RC formation *in vivo* in metazoan.

Such information, coupled with origin activity and sequence information could be interesting to further identify the diversity of origins and their dynamics.

Nucleosome positioning over replication origins

From our minimal β^A-globin origin, we could investigate nucleosomes positioning on an active and non-active origin, which gave us a causal link of nucleosomes landscape with replication origin activity. Although the strongest hypothesis is that we detect nucleosomes, (sizes of the fragments correspond to mono-nucleosomes and the MNase digestion is important) it remains to confirm the nature of protein protecting DNA from the MNase digestion. One way is to use MNase-ChIP-Seq, once digested by the MNase[109], the sample is suggested to a ChIP, using antibodies directed against core histones proteins. Recovered peaks will confirm nucleosome positioning. It would also be interesting to investigate the presence of the histone variant H2A.Z since it was proposed to play a crucial role in origin licensing[125]. Comparison between the β^A-globin minimal origin and the inactive one containing pG4#1 on the opposite strand could reveal differences of the nucleosome landscape that could help to understand the mechanisms of DNA replication initiation. In mammals, the presence of labile nucleosomes over replication origins and replication initiation sites has been described. This observation was made thanks to different cross-linking and MNase digestions[122]. Since we did not address this question, we cannot conclude on the presence or not of a labile nucleosome over our minimal origin but it would be of great interest to answer this question.

Nucleosomes positioning over the β^A-globin minimal origins using MNase-seq was made by sequencing genome-wide to also recover nucleosomes positioning information from the genome. This genome-wide approach was made in cells synchronised at the G1/S transition and in G2/M. Since we identified

a specific pattern over our active origins, it would be of great interest to investigate the nucleosome positioning over active origins found all over the genome. With our already made SNS map, we could focus on the nucleosome dynamics over replication origins. We could reach a deep analysis similar to that already made in budding yeast, but also in mammals[116,118,122]. However, so far no nucleosome mapping at the two key steps of the cell cycle were compared with a map of origins in vertebrates genome-wide. The presence of phased nucleosomes at replication origins and their dynamics (the +3 nucleosomes of the β^A-globin minimal origin (Figure 6 from the results) should be investigated to generalise this observation to active replication origins.

One limitation of such approach could be the lack of a characteristic nucleosomes landscape on replication origins localised inside promoters. Indeed, as already discussed, it is possible that replication origin, inside a promoter, does not affect the nucleosomes landscape that depends mostly on transcription *cis*-regulatory elements.

Bibliography

1. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
2. Alloway, J. L. FURTHER OBSERVATIONS ON THE USE OF PNEUMOCOCCUS EXTRACTS IN EFFECTING TRANSFORMATION OF TYPE IN VITRO. *J. Exp. Med.* **57**, 265–278 (1933).
3. Avery, O. T., MacLeod, C. M. & McCarty, M. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES. *J. Exp. Med.* **79**, 137–158 (1944).
4. Meselson, M. & Stahl, F. W. THE REPLICATION OF DNA IN ESCHERICHIA COLI*. *Proc. Natl. Acad. Sci. U. S. A.* **44**, 671–682 (1958).
5. Jacob, F. & Brenner, S. [On the regulation of DNA synthesis in bacteria: the hypothesis of the replicon]. *Comptes Rendus Hebd. Seances Acad. Sci.* **256**, 298–300 (1963).
6. Barnum, K. J. & O'Connell, M. J. Cell Cycle Regulation by Checkpoints. *Methods Mol. Biol. Clifton NJ* **1170**, 29–40 (2014).
7. Cutter, A. & Hayes, J. J. A Brief Review of Nucleosome Structure. *FEBS Lett.* **589**, 2914–2922 (2015).
8. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
9. Zhang, Y. *et al.* Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nat. Struct. Mol. Biol.* **16**, 847–852 (2009).
10. Chiarella, A. M., Lu, D. & Hathaway, N. A. Epigenetic Control of a Local Chromatin Landscape. *Int. J. Mol. Sci.* **21**, 943 (2020).
11. Biterge, B. & Schneider, R. Histone variants: key players of chromatin. *Cell Tissue Res.* **356**, 457–466 (2014).
12. Jin, C. *et al.* H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions in the human genome. *Nat. Genet.* **41**, 941–945 (2009).

13. Fragkos, M., Ganier, O., Coulombe, P. & Méchali, M. DNA replication origin activation in space and time. *Nat. Rev. Mol. Cell Biol.* **16**, 360–374 (2015).
14. Bell, S. P. & Stillman, B. ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. *Nature* **357**, 128–134 (1992).
15. Bowers, J. L., Randell, J. C. W., Chen, S. & Bell, S. P. ATP Hydrolysis by ORC Catalyzes Reiterative Mcm2-7 Assembly at a Defined Origin of Replication. *Mol. Cell* **16**, 967–978 (2004).
16. Tanaka, S. & Diffley, J. F. X. Interdependent nuclear accumulation of budding yeast Cdt1 and Mcm2–7 during G1 phase. *Nat. Cell Biol.* **4**, 198–207 (2002).
17. Speck, C. & Stillman, B. Cdc6 ATPase ACTIVITY REGULATES ORC-Cdc6 STABILITY AND THE SELECTION OF SPECIFIC DNA SEQUENCES AS ORIGINS OF DNA REPLICATION. *J. Biol. Chem.* **282**, 11705–11714 (2007).
18. McGarry, T. J. & Kirschner, M. W. Geminin, an Inhibitor of DNA Replication, Is Degraded during Mitosis. *Cell* **93**, 1043–1053 (1998).
19. Fernández-Cid, A. *et al.* An ORC/Cdc6/MCM2-7 Complex Is Formed in a Multistep Reaction to Serve as a Platform for MCM Double-Hexamer Assembly. *Mol. Cell* **50**, 577–588 (2013).
20. Frigola, J., Remus, D., Mehanna, A. & Diffley, J. F. X. ATPase-dependent quality control of DNA replication origin licensing. *Nature* **495**, 339–343 (2013).
21. 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**, 704–710 (2019).
22. Parker, M. W., Botchan, M. R. & Berger, J. M. Mechanisms and regulation of DNA replication initiation in eukaryotes. *Crit. Rev. Biochem. Mol. Biol.* **52**, 107–144 (2017).
23. Yeeles, J. T. P., Deegan, T. D., Janska, A., Early, A. & Diffley, J. F. X. Regulated Eukaryotic DNA Replication Origin Firing with Purified Proteins. *Nature* **519**, 431–435 (2015).
24. Renard-Guillet, C., Kanoh, Y., Shirahige, K. & Masai, H. Temporal and spatial regulation of eukaryotic DNA replication: From regulated initiation to genome-scale timing program. *Semin. Cell Dev. Biol.* **30**, 110–120 (2014).

25. Zhao, P. A., Sasaki, T. & Gilbert, D. M. High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome Biol.* **21**, (2020).
26. Hiratani, I. *et al.* Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol.* **6**, e245 (2008).
27. Ryba, T. *et al.* Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.* **20**, 761–770 (2010).
28. Pope, B. D. *et al.* Topologically-associating domains are stable units of replication-timing regulation. *Nature* **515**, 402–405 (2014).
29. Marchal, C., Sima, J. & Gilbert, D. M. Control of DNA replication timing in the 3D genome. *Nat. Rev. Mol. Cell Biol.* **20**, 721–737 (2019).
30. Lin, C. M., Fu, H., Martinovsky, M., Bouhassira, E. & Aladjem, M. I. Dynamic Alterations of Replication Timing in Mammalian Cells. *Curr. Biol.* **13**, 1019–1028 (2003).
31. Hassan-Zadeh, V. *et al.* USF Binding Sequences from the HS4 Insulator Element Impose Early Replication Timing on a Vertebrate Replicator. *PLoS Biol.* **10**, (2012).
32. Blin, M. *et al.* Transcription-dependent regulation of replication dynamics modulates genome stability. *Nat. Struct. Mol. Biol.* **26**, 58 (2019).
33. Brueckner, L. *et al.* Local rewiring of genome–nuclear lamina interactions by transcription. *EMBO J.* **39**, (2020).
34. Sima, J. *et al.* Identifying cis Elements for Spatiotemporal Control of Mammalian DNA Replication. *Cell* **176**, 816-830.e18 (2019).
35. Hiraga, S. *et al.* Human RIF1 and protein phosphatase 1 stimulate DNA replication origin licensing but suppress origin activation. *EMBO Rep.* **18**, 403–419 (2017).
36. Hiraga, S. *et al.* Rif1 controls DNA replication by directing Protein Phosphatase 1 to reverse Cdc7-mediated phosphorylation of the MCM complex. *Genes Dev.* **28**, 372–383 (2014).

37. Foti, R. *et al.* Nuclear Architecture Organized by Rif1 Underpins the Replication-Timing Program. *Mol. Cell* **61**, 260–273 (2016).

38. Knott, S. R. V. *et al.* Forkhead Transcription Factors Establish Origin Timing and Long-Range Clustering in S. cerevisiae. *Cell* **148**, 99–111 (2012).

39. 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).

40. Zhang, H. *et al.* Dynamic relocalization of replication origins by Fkh1 requires execution of DDK function and Cdc45 loading at origins. *eLife* **8**,.

41. Dellino, G. I. *et al.* Genome-wide mapping of human DNA-replication origins: Levels of transcription at ORC1 sites regulate origin selection and replication timing. *Genome Res.* **23**, 1–11 (2013).

42. Miotto, B., Ji, Z. & Struhl, K. Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers. *Proc. Natl. Acad. Sci.* **113**, E4810–E4819 (2016).

43. Sugimoto, N., Maehara, K., Yoshida, K., Ohkawa, Y. & Fujita, M. Genome-wide analysis of the spatiotemporal regulation of firing and dormant replication origins in human cells. *Nucleic Acids Res.* doi:10.1093/nar/gky476.

44. Powell, S. K. *et al.* Dynamic loading and redistribution of the Mcm2-7 helicase complex through the cell cycle. *EMBO J.* **34**, 531–543 (2015).

45. Gros, J. *et al.* Post-licensing specification of eukaryotic replication origins by facilitated Mcm2-7 sliding along DNA. *Mol. Cell* **60**, 797–807 (2015).

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

47. Smith, D. J. & Whitehouse, I. Intrinsic coupling of lagging-strand synthesis to chromatin assembly. *Nature* **483**, 434–438 (2012).

48. Petryk, N. *et al.* Replication landscape of the human genome. *Nat. Commun.* **7**, (2016).

49. McGuffee, S. R., Smith, D. J. & Whitehouse, I. Quantitative, Genome-Wide Analysis of Eukaryotic Replication Initiation and Termination. *Mol. Cell* **50**, 123–135 (2013).

50. Chen, Y.-H. *et al.* Transcription shapes DNA replication initiation and termination in human cells. *Nat. Struct. Mol. Biol.* **26**, 67–77 (2019).

51. Prioleau, M.-N., Gendron, M.-C. & Hyrien, O. Replication of the chicken beta-globin locus: early-firing origins at the 5' HS4 insulator and the rho- and betaA-globin genes show opposite epigenetic modifications. *Mol. Cell. Biol.* **23**, 3536–3549 (2003).

52. Cadoret, J.-C. *et al.* Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 15837–15842 (2008).

53. Sequeira-Mendes, J. *et al.* Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genet.* **5**, e1000446 (2009).

54. 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).

55. Foulk, M. S., Urban, J. M., Casella, C. & Gerbi, S. A. Characterizing and controlling intrinsic biases of lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-quadruplex motifs around a subset of human replication origins. *Genome Res.* **25**, 725–735 (2015).

56. Picard, F. *et al.* The Spatiotemporal Program of DNA Replication Is Associated with Specific Combinations of Chromatin Marks in Human Cells. *PLoS Genet.* **10**, (2014).

57. DePamphilis, M. L. EUKARYOTIC DNA REPLICATION: Anatomy of An Origin. 37.

58. Stinchcomb, D. T., Struhl, K. & Davis, R. W. Isolation and characterisation of a yeast chromosomal replicator. *Nature* **282**, 39–43 (1979).

59. Kearsey, S. Analysis of sequences conferring autonomous replication in baker's yeast. *EMBO J.* **2**, 1571–1575 (1983).

60. Celniker, S. E., Sweder, K., Srienc, F., Bailey, J. E. & Campbell, J. L. Deletion mutations affecting autonomously replicating sequence ARS1 of Saccharomyces cerevisiae. *Mol. Cell. Biol.* **4**, 2455–2466 (1984).

61. Diffley, J. F. X. & Cocker, J. H. Protein-DNA interactions at a yeast replication origin. *Nature* **357**, 169–172 (1992).

62. Wyrick, J. J. *et al.* Genome-Wide Distribution of ORC and MCM Proteins in S. cerevisiae: High-Resolution Mapping of Replication Origins. *Science* **294**, 2357–2360 (2001).

63. Breier, A. M., Chatterji, S. & Cozzarelli, N. R. Prediction of Saccharomyces cerevisiae replication origins. *Genome Biol.* **5**, R22 (2004).

64. Singh, V. K. & Krishnamachari, A. Context based computational analysis and characterization of ARS consensus sequences (ACS) of Saccharomyces cerevisiae genome. *Genomics Data* **9**, 130–136 (2016).

65. Coster, G. & Diffley, J. F. X. Bidirectional eukaryotic DNA replication is established by quasi-symmetrical helicase loading. *Science* **357**, 314–318 (2017).

66. Comoglio, F. *et al.* High-resolution profiling of Drosophila replication start sites reveals a DNA shape and chromatin signature of metazoan origins. *Cell Rep.* **11**, 821–834 (2015).

67. Besnard, E. *et al.* Unraveling cell type–specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat. Struct. Mol. Biol.* **19**, 837–844 (2012).

68. Cayrou, C. *et al.* New insights into replication origin characteristics in metazoans. *Cell Cycle* **11**, 658–667 (2012).

69. Massip, F. *et al.* Evolution of replication origins in vertebrate genomes: rapid turnover despite selective constraints. *Nucleic Acids Res.* **47**, 5114–5125 (2019).

70. Cayrou, C. *et al.* The chromatin environment shapes DNA replication origin organization and defines origin classes. *Genome Res.* **25**, 1873–1885 (2015).

71. Aladjem, M. I., Rodewald, L. W., Kolman, J. L. & Wahl, G. M. Genetic Dissection of a Mammalian Replicator in the Human β-Globin Locus. *Science* **281**, 1005–1009 (1998).

72. Valton, A.-L. *et al.* G4 motifs affect origin positioning and efficiency in two vertebrate replicators. *EMBO J.* **33**, 732–746 (2014).

73. Prorok, P. *et al.* Involvement of G-quadruplex regions in mammalian replication origin activity. *Nat. Commun.* **10**, (2019).

74. Gellert, M., Lipsett, M. N. & Davies, D. R. HELIX FORMATION BY GUANYLIC ACID. *Proc. Natl. Acad. Sci. U. S. A.* **48**, 2013–2018 (1962).

75. Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K. & Neidle, S. Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.* **34**, 5402–5415 (2006).

76. Rhodes, D. & Lipps, H. J. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res.* **43**, 8627–8637 (2015).

77. Bugaut, A. & Balasubramanian, S. A Sequence-Independent Study of the Influence of Short Loop Lengths on the Stability and Topology of Intramolecular DNA G-Quadruplexes. *Biochemistry* **47**, 689–697 (2008).

78. Guédin, A., Gros, J., Alberti, P. & Mergny, J.-L. How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic Acids Res.* **38**, 7858–7868 (2010).

79. Mukundan, V. T. & Phan, A. T. Bulges in G-Quadruplexes: Broadening the Definition of G-Quadruplex-Forming Sequences. *J. Am. Chem. Soc.* **135**, 5017–5028 (2013).

80. Fernando, H., Rodriguez, R. & Balasubramanian, S. Selective Recognition of a DNA G-Quadruplex by an Engineered Antibody. *Biochemistry* **47**, 9365–9371 (2008).

81. Lam, E. Y. N., Beraldi, D., Tannahill, D. & Balasubramanian, S. G-Quadruplex structures are stable and detectable in human genomic DNA. *Nat. Commun.* **4**, 1796 (2013).

82. Huppert, J. L. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.* **33**, 2908–2916 (2005).

83. Biffi, G., Tannahill, D., McCafferty, J. & Balasubramanian, S. Quantitative Visualization of DNA G-quadruplex Structures in Human Cells. *Nat. Chem.* **5**, 182–186 (2013).

84. Chambers, V. S. *et al.* High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.* **33**, 877–881 (2015).

85. Hänsel-Hertsch, R. *et al.* G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.* **48**, 1267–1272 (2016).

86. Huppert, J. L. & Balasubramanian, S. G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.* **35**, 406–413 (2007).

87. Maizels, N. & Gray, L. T. The G4 Genome. *PLOS Genet.* **9**, e1003468 (2013).

88. Ribeyre, C. *et al.* The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. *PLoS Genet.* **5**, e1000475 (2009).

89. Lopes, J. *et al.* G-quadruplex-induced instability during leading-strand replication. *EMBO J.* **30**, 4033–4046 (2011).

90. Piazza, A. *et al.* Short loop length and high thermal stability determine genomic instability induced by G-quadruplex-forming minisatellites. *EMBO J.* **34**, 1718–1734 (2015).

91. Sarkies, P., Reams, C., Simpson, L. J. & Sale, J. E. Epigenetic Instability due to Defective Replication of Structured DNA. *Mol. Cell* **40**, 703–713 (2010).

92. Sarkies, P. *et al.* FANCJ coordinates two pathways that maintain epigenetic stability at G-quadruplex DNA. *Nucleic Acids Res.* **40**, 1485–1498 (2012).

93. Schiavone, D. *et al.* Determinants of G quadruplex-induced epigenetic instability in REV1-deficient cells. *EMBO J.* **33**, 2507–2520 (2014).

94. McRae, E. K. S., Booy, E. P., Padilla-Meier, G. P. & McKenna, S. A. On Characterizing the Interactions between Proteins and Guanine Quadruplex Structures of Nucleic Acids. *J. Nucleic Acids* **2017**, (2017).

95. González, V., Guo, K., Hurley, L. & Sun, D. Identification and Characterization of Nucleolin as a c-myc G-quadruplex-binding Protein. *J. Biol. Chem.* **284**, 23622–23635 (2009).

96. Lago, S., Tosoni, E., Nadai, M., Palumbo, M. & Richter, S. N. The cellular protein nucleolin preferentially binds long-looped G-quadruplex nucleic acids. *Biochim. Biophys. Acta* **1861**, 1371–1381 (2017).

97. Kanoh, Y. *et al.* Rif1 binds to G quadruplexes and suppresses replication over long distances. *Nat. Struct. Mol. Biol.* **22**, 889–897 (2015).

98. Kumagai, A. & Dunphy, W. G. MTBP, the partner of Treslin, contains a novel DNA-binding domain that is essential for proper initiation of DNA replication. *Mol. Biol. Cell* **28**, 2998–3012 (2017).

99. Keller, H. *et al.* The intrinsically disordered amino-terminal region of human RecQL4: multiple DNA-binding domains confer annealing, strand exchange and G4 DNA binding. *Nucleic Acids Res.* **42**, 12614–12627 (2014).

100. Hoshina, S. *et al.* Human Origin Recognition Complex Binds Preferentially to G-quadruplex-preferable RNA and Single-stranded DNA. *J. Biol. Chem.* **288**, 30161–30171 (2013).

101. Lai, W. K. M. & Pugh, B. F. Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nat. Rev. Mol. Cell Biol.* **18**, 548–562 (2017).

102. Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning11Edited by T. Richmond. *J. Mol. Biol.* **276**, 19–42 (1998).

103. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).

104. Arya, G., Maitra, A. & Grigoryev, S. A. A Structural Perspective on the Where, How, Why, and What of Nucleosome Positioning. *J. Biomol. Struct. Dyn.* **27**, 803–820 (2010).

105. Cairns, B. R. The logic of chromatin architecture and remodelling at promoters. *Nature* **461**, 193–198 (2009).

106. Giaimo, B. D., Ferrante, F., Herchenröther, A., Hake, S. B. & Borggrefe, T. The histone variant H2A.Z in gene regulation. *Epigenetics Chromatin* **12**, (2019).

107. Vasseur, P. *et al.* Dynamics of Nucleosome Positioning Maturation following Genomic Replication. *Cell Rep.* **16**, 2651–2665 (2016).

108. Chereji, R. V., Bryson, T. D. & Henikoff, S. Quantitative MNase-seq accurately maps nucleosome occupancy levels. *Genome Biol.* **20**, 198 (2019).

109. Chereji, R. V. *et al.* Genome-wide profiling of nucleosome sensitivity and chromatin accessibility in Drosophila melanogaster. *Nucleic Acids Res.* **44**, 1036–1051 (2016).

110. Chereji, R. V., Ocampo, J. & Clark, D. J. MNase-Sensitive Complexes in Yeast: Nucleosomes and Non-histone Barriers. *Mol. Cell* **65**, 565-577.e3 (2017).

111. Allan, J., Fraser, R. M., Owen-Hughes, T. & Keszenman-Pereyra, D. Micrococcal Nuclease Does Not Substantially Bias Nucleosome Mapping. *J. Mol. Biol.* **417–135**, 152–164 (2012).

112. Brogaard, K., Xi, L., Wang, J.-P. & Widom, J. A base pair resolution map of nucleosome positions in yeast. *Nature* **486**, 496–501 (2012).

113. Chereji, R. V., Ramachandran, S., Bryson, T. D. & Henikoff, S. Precise genome-wide mapping of single nucleosomes and linkers in vivo. *Genome Biol.* **19**, 19 (2018).

114. Simpson, R. T. Nucleosome positioning can affect the function of a cis- acting DNA element in vivo. 3 (1990).

115. Lipford, J. R. & Bell, S. P. Nucleosomes Positioned by ORC Facilitate the Initiation of DNA Replication. *Mol. Cell* **7**, 21–30 (2001).

116. Eaton, M. L., Galani, K., Kang, S., Bell, S. P. & MacAlpine, D. M. Conserved nucleosome positioning defines replication origins. *Genes Dev.* **24**, 748–753 (2010).

117. Berbenetz, N. M., Nislow, C. & Brown, G. W. Diversity of Eukaryotic DNA Replication Origins Revealed by Genome-Wide Analysis of Chromatin Structure. *PLOS Genet.* **6**, e1001092 (2010).

118. 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**, 212–224 (2015).

119. Rodriguez, J., Lee, L., Lynch, B. & Tsukiyama, T. Nucleosome occupancy as a novel chromatin parameter for replication origin functions. *Genome Res.* **27**, 269–277 (2017).

120. MacAlpine, H. K., Gordân, R., Powell, S. K., Hartemink, A. J. & MacAlpine, D. M. Drosophila ORC localizes to open chromatin and marks sites of cohesin complex loading. *Genome Res.* **20**, 201–211 (2010).

121. Lubelsky, Y. *et al.* Pre-replication complex proteins assemble at regions of low nucleosome occupancy within the Chinese hamster dihydrofolate reductase initiation zone. *Nucleic Acids Res.* **39**, 3141–3155 (2011).

122. Lombraña, R. *et al.* High-resolution analysis of DNA synthesis start sites and nucleosome architecture at efficient mammalian replication origins. *EMBO J.* **32**, 2631–2644 (2013).

123. Deal, R. B., Henikoff, J. G. & Henikoff, S. Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones. *Science* **328**, 1161–1164 (2010).

124. Zentner, G. E., Kasinathan, S., Xin, B., Rohs, R. & Henikoff, S. ChEC-seq kinetics discriminates transcription factor binding sites by DNA sequence and shape in vivo. *Nat. Commun.* **6**, (2015).

125. Long, H. *et al.* H2A.Z facilitates licensing and activation of early replication origins. *Nature* (2019) doi:10.1038/s41586-019-1877-9.

La réplication de l'ADN est un mécanisme critique qui se déroule lors du cycle cellulaire. Sous contrôle d'un programme spatio-temporel, la réplication démarre aux niveaux des origines de réplication. L'identification de ces origines chez l'humain et la souris a mis en évidence des motifs G-quadruplexes (pG4). Des analyses génétiques ont montré leur rôle essentiel dans l'activité des origines de réplications. Cependant les G4s ne sont pas suffisant pour définir une origine de réplication, ils doivent être associés à d'autre éléments *cis*-régulateurs qui n'ont pas encore été identifiés. Afin de mettre en évidence de nouveaux éléments essentiels à l'activité origine, nous avons utilisé les origines modèles β^A-globine et Med14. L'étude de ces origines a permis de mettre en avant la nécessité d'un second pG4, ainsi que la participation des boites CCAAT et TATA. Il est devenu alors possible d'obtenir une séquence minimale, de 90 pb, capable d'induire une initiation de la réplication très efficace. De plus, la boite TATA est essentielle au contrôle du moment de réplication. Cette origine constitue un outil moléculaire sans précédent qui permet l'étude mécanistique des origines de réplication chez les vertébrés. En effet, grâce à cette origine minimale, nous avons pu montrer que les deux pG4s présents dans la séquence devaient être situés sur le même brin afin d'aboutir à une activité origine. Une étude du positionnement des nucléosomes sur cette origine minimale a mis en évidence la présence d'une région dépourvue de nucléosomes au niveau des deux pG4s, séparée du site d'initiation par un nucléosome fortement positionné. Ces résultats sont en accord avec ceux déjà trouvés à l'échelle du génome, confirmant et validant notre analyse et notre origine minimale modèle. De plus, notre étude amène de nouveaux éléments quant à la définition des origines de réplications.

Mots clefs : Origines de réplication, G-quadruplexes, positionnement de nucléosome, réplication.

DNA replication is a critical mechanism that occur during the cell cycle. DNA replication is under the control of a spatiotemporal program that ensures the accurate duplication of the genome. Many genome-wide studies in vertebrates identified promoters containing G-quadruplex motifs (pG4s) as being strongly associated with efficient replication origins. Their necessity has been proven genetically. However, it was shown that one pG4 was not sufficient. We define here, a minimal 90 bp model origin containing two pG4s located on the same strand and associated with a CCAAT and a TATA-boxes. The TATA-box participates in the replication timing control of this minimal origin. The transfer of one pG4 on the other strand fully abolishes origin function but the distance between them is quite flexible. Analyses of nucleosome organisation over the minimal origin reveals a very specific organisation previously found genome-wide in vertebrates. The origin is inside a NDR and separated by the site of initiation by a strongly positioned nucleosome. In conclusion, our study provides a new paradigm for the genetic and nucleosome organisation of vertebrate origins.

Key words: Replication origins, G-quadruplex, nucleosome positioning, replication.