Chromatin-dependent pre-replication complex positioning and activation in mammals

Nina Danielle Kirstein

To cite this version:

Nina Danielle Kirstein. Chromatin-dependent pre-replication complex positioning and activation in mammals. Human health and pathology. Université Montpellier; Ludwig-Maximilians Universität (Munich, Allemagne), 2017. English. NNT : 2017MONTT005 . tel-01557752

THÈSE

Pour obtenir le grade de
Docteur

Délivré par **l'Université de Montpellier et l'Université Ludwig Maximilian de Munich**

Préparée au sein de l'école doctorale CBS2
Et de l'unité de recherche *Institute of Regenerative Medicine and Biotherapy* et *Helmholtz Zentrum München*

Spécialité : **Biologie/ Santé**

Présentée par **Nina Danielle KIRSTEIN**

Chromatin-Dependent Pre-Replication Complex Positioning and Activation in Mammals

Soutenue le 8. Juin 2017 devant le jury composé de

Dr. María GOMEZ, Centro de Biologia Molecular (ES)	Rapporteur
Dr. Torsten KRUDE, Univ. of Cambridge (UK)	Rapporteur
Prof. Dr. Axel IMHOF, Univ. Ludwig Maximilian (DE)	Président du jury
Dr. Philippe PASERO, IGH (FR)	Examinateur
Prof. Dr. Gunnar SCHOTTA, Univ. Ludwig Maximilian (DE)	Examinateur
Prof. Dr. John DE VOS, IRMB (FR)	Examinateur
Dr. Jean-Marc LEMAITRE, IRMB (FR)	Codirecteur de thèse
Dr. Aloys SCHEPERS, Helmholtz Zentrum München (DE)	Directeur de thèse

RESUME

Chaque division cellulaire requiert une duplication précise du génome. Des dizaines de milliers de sites d'initiation de la réplication d'ADN (origines de réplication) sont impliqués dans la réplication complète du génome humain. L'activation des origines de réplication est régulée précisément et des études génomiques extensives ont démontré la présence de caractéristiques génomiques associées à l'activation des origines de réplication. Le complexe de pré-réplication (pre-RC) est la base de l'initiation de la réplication et consiste en deux sous-complexes majeurs : l' « origin recognition complex » (ORC) qui interagit directement avec l'ADN et est nécessaire pour recruter le second sous-complexe, les hélicases Mcm2-7, qui sont responsables de l'initiation de la réplication. La régulation de l'assemblage du pre-RC est bien étudiée, mais les caractéristiques de la chromatine qui déterminent le positionnement du pre-RC sur le génome restent peu connues. Les études génomiques par immuno-précipitation de la chromatine et séquençage à haut débit (ChIP-seq) des pre-RCs sont rares et jusqu'à aujourd'hui seulement disponibles pour ORC. Du fait que Mcm2-7 migre de son site de chargement initial, il est crucial d'obtenir des informations sur le positionnement des Mcm2-7 pour la compréhension complète de la régulation de la réplication.

Ce travail présente la première analyse génomique par méthode ChIP-seq des deux sous-unités majeures du pre-RC, ORC et Mcm2-7, dans la lignée cellulaire de lymphome de Burkitt Raji infectée par le virus d'Epstein-Barr (EBV). La présence du génome d'EBV permet d'avoir un contrôle interne de la qualité de nos expériences, en comparant les positions de pre-RC déterminées avec des positions du pre-RC précédemment publiées. Sur le génome humain, les résultats de séquençage du pre-RC corrèlent bien avec des zones de réplication active. De façon intéressante, les zones de terminaison de la réplication étaient spécifiquement bas en pre-RC, spécialement en Mcm2-7. La localisation des sites d'initiation de la réplication identifiés est généralement bien corrélée avec les sites de transcription active. En effet, des sites d'assemblage du pre-RC de haute affinité sont localisés préférentiellement en voisinage de sites de transcription active, ce qui est possiblement dû à l'accessibilité de la chromatine dans ces régions. La fixation de Mcm2-7 fluctuait de façon dépendante du cycle cellulaire, ce qui suggère des translocations de Mcm2-7 en G1, probablement dépendantes de la machinerie active de la transcription. Ces résultats indiquent que les positions de ORC et Mcm2-7 sont principalement dépendantes de l'accessibilité de la chromatine avec un accès privilégié dans la chromatine active et Mcm2-7 étant le déterminant majeur de l'initiation de la réplication.

Au sein de l'hétérochromatine, ORC était enrichi dans des zones associées avec l'histone modifié H4K20me3. Cependant, cet enrichissement était moins important pour les Mcm2-7. En utilisant un système de réplication basé sur des plasmides, nous avons démontré que l'association d'ORC et H4K20me3 favorise l'assemblage du pre-RC et l'initiation de la réplication. Cette observation suggère que l'interaction ORC-chromatine est le déterminant majeur de la régulation de la réplication d'ADN au sein de l'hétérochromatine. En conclusion, cette étude propose deux mécanismes différents de la régulation de l'assemblage du pre-RC dépendants de l'environnement de la chromatine.

Mots clés : Réplication d'ADN, pre-RC, chromatine, PR-Set7, Suv4-20h1/h2, H4K20 méthylation

ZUSAMMENFASSUNG

Mit jeder Zellteilung muss das Genom exakt dupliziert werden. Zehntausende Replikationsinitiationsstellen sind bei der Replikation des gesamten humanen Genoms beteiligt. Die Aktivierung der Initiationsstellen ist präzise reguliert und umfassende Genom-weite Studien haben verschiedene genomische Faktoren identifiziert, die die Aktivierung der Replikationsinitiationsstellen beeinflussen. Der Prä-Replikationskomplex (pre-RC) bildet die Grundlage der Replikationsinitiation und besteht aus zwei Hauptuntereinheiten: der „origin recognition complex“ (ORC) bindet DNS und wird zum Laden der zweiten Untereinheit, den Mcm2-7 Helikasen benötigt, die die eigentliche Replikationsinitiation veranlassen. Während die Regulation des pre-RC Aufbaus vielfach untersucht wurde und mittlerweile gut verstanden wird, sind die Chromatinkomponenten, welche die Positionierung der pre-RCs regulieren, weitgehend unbekannt. Die wenigen Genom-weiten pre-RC Chromatin Immunopräzipitations- und Sequenzierungsstudien (ChIP-seq), behandeln bis heute nur ORC. Da sich Mcm2-7 allerdings von seiner initialen Ladeposition fortbewegen kann, werden vor allem die Genom-weiten Positionen von Mcm2-7 benötigt, um die Regulation der DNS Replikation vollständig zu verstehen.

Diese Arbeit umfasst die erste Genom-weite pre-RC ChIP-seq Analyse der zwei pre-RC Hauptkomponenten ORC und Mcm2-7 in der Epstein-Barr Virus (EBV) infizierten Burkitt-Lymphom Zelllinie Raji. Als Qualitätskontrolle für erfolgreiche ChIPs wurden die aus den vorliegenden Experimenten bestimmten pre-RC Positionen auf dem EBV-Genom mit bereits bekannten pre-RC Positionen verglichen. Auf dem humanen Genom korrelierten die pre-RC ChIP Ergebnisse mit aktiven Replikationszonen, während Replikationsterminationszonen eine spezifische Abnahme der pre-RC Komponenten, besonders von Mcm2-7, aufzeigten. Es ist bereits bekannt, dass aktive Replikation mit aktiver Transkription korreliert. Starke pre-RC Bindung war in der Tat hauptsächlich an Regulationsstellen der aktiven Transkription zu finden, was vermutlich durch die Zugänglichkeit des Chromatins determiniert wird. Starke Mcm2-7 Bindung variierte dabei in Abhängigkeit des Zellzyklus, was für Mcm2-7 Translokationen während der G1 Phase spricht, die vermutlich von der aktiven Transkriptionsmaschinerie beeinflusst werden. Diese Ergebnisse deuten darauf hin, dass ORC und Mcm2-7 Positionen in der aktiven Chromatinumgebung hauptsächlich von der Zugänglichkeit des Chromatins abhängen und Mcm2-7 die Hauptkomponente der Bestimmung der Replikationsinitiationsstellen darstellt.

In Heterochromatin assoziierte vorwiegend ORC mit der heterochromatischen Histonmodifikation H4K20me3, während Mcm2-7 weniger Anreicherung zeigte. Unter Verwendung eines Plasmid-basierten Replikationssystems wurde bestätigt, dass diese ORC-Chromatin Assoziation einen essentiellen Einfluss auf die Regulation der pre-RC Positionierung und Aktivierung ausübt. Dieses Ergebnis zeigt, dass ORC-Chromatin Interaktionen einen entscheidenden Faktor für die Regulation der Replikation in Heterochromatin darstellen. Zusammenfassed schlägt diese Studie zwei verschiedene Modi der pre-RC Positionierung vor, welche jeweils vom Chromatinkontext abhängen.

Schlagwörter : DNA Replikation, pre-RC, Chromatin, PR-Set7, Suv4-20h1/h2, H4K20 Methylierung

ABSTRACT

With every cell division, the genome needs to be faithfully duplicated. Tens of thousands of DNA replication initiation sites (origins of replication) are involved in replicating the human genome. Origin activation is precisely regulated and extensive genome-wide studies found association of origin activation to several different genomic features. The pre-replication complex (pre-RC) is the basis for replication initiation and consists of two major subcomponents: the origin recognition complex (ORC) binds DNA and is required for loading of the second component, Mcm2-7 helicases, which initiate DNA replication. Regulation of pre-RC assembly is well studied, however, chromatin features driving pre-RC positioning on the human genome remain largely unknown. Genome-wide pre-RC chromatin immunoprecipitation experiments followed by sequencing (ChIP-seq) studies are rare and so far only performed for ORC. As Mcm2-7 can translocate from their initial loading site, information about Mcm2-7 positioning are required for full understanding of DNA replication regulation.

This work presents the first genome-wide ChIP-seq analysis of the two major pre-RC subcomponents ORC and Mcm2-7 in the Epstein-Barr virus (EBV) infected Burkitt's lymphoma cell line Raji. Successful ChIPs were validated on the EBV genome by comparing obtained pre-RC positions with already existing pre-RC ChIP-on chip data. On the human genome, pre-RC sequencing results nicely correlated with zones of active replication. Interestingly, zones of replication termination were specifically depleted from pre-RC components, especially from Mcm2-7. Active DNA replication is known to correlate with active transcription. Indeed, strong pre-RC assembly preferentially occurred at sites of active transcriptional regulation, presumably determined by chromatin accessibility. Strong Mcm2-7 binding thereby fluctuated cell cycle-dependently, arguing for Mcm2-7 translocations during G1, possibly depending on the active transcriptional machinery. These results indicate ORC and Mcm2-7 positions being mainly dependent on chromatin accessibility in active chromatin, with Mcm2-7 being the major determinant of replication initiation.

In heterochromatin, ORC was enriched at H4K20me3 sites, while Mcm2-7 enrichment was less prominent. Employing a plasmid-based replication system, ORC association to H4K20me3 was proven to promote successful pre-RC assembly and replication initiation, situating direct ORC-chromatin interactions being the major determinant for DNA replication regulation in heterochromatin. Taken together, this study proposes two different modes of pre-RC assembly regulation depending on chromatin environment.

Keywords: DNA replication, origin licensing, pre-RC, chromatin, PR-Set7, Suv4-20h1/h2, H4K20methylation

TABLE OF CONTENTS

1. INTRODUCTION

DNA replication is the process of precisely copying the complete genetic information, which assures faithful inheritance of the genome during each cell division. Thereby, DNA replication initiates only once per cell cycle. Misregulation during this process leads to genetic instability, which might have fatal consequences on organs and tissues and can cause cancer or other genetic disorders in humans (Abbas, Keaton, and Dutta 2013). Sites of DNA replication initiation are called origins of replication. While bacteria initiate genome replication at one specific origin (Mott and Berger 2007), replication regulation in *Saccharomyces cerevisiae* is already more complex. In *S. cerevisiae*, DNA replication origins were initially identified as autonomous replication sequences (ARS), which contain an 11-17 bp AT-rich consensus sequence. However, only a small part of all ARS consensus sequences are used as replication origins, suggesting that other features also contribute to origin recognition and activation (Fragkos *et al.* 2015). In higher eukaryotes, origins do not exhibit sequence specificities and DNA replication can initiate at any DNA sequence (Vashee *et al.* 2003). In humans, 30000-50000 origins are activated in each cell at each cell cycle. Numerous studies during the recent years revealed that origins exhibit a preference for specific features, although none of these features *alone* are predictive for replication origins (Méchali 2010). In the following, I will detail the regulation of DNA replication in higher eukaryotes and the impact of chromatin, histone modifications, and structural arrangements within the nucleus.

1.1 DNA REPLICATION IS TIGHTLY REGULATED – AND STOCHASTIC

DNA replication is spatio-temporally separated to ensure correct duplication of the genome. During each S-phase of the cell cycle, 30000-50000 origins are activated per cell. These origins are chosen from an exceeding pool of possible origins. For an origin to be activated, proper loading of the pre-replication complex (pre-RC) during preceding G1 is required.

1.1.1 ASSEMBLY OF THE PRE-REPLICATION COMPLEX IN LATE MITOSIS/ EARLY G1

The process of pre-RC assembly is called origin licensing and is restricted from late mitosis to the restriction point in G1-phase of the cell cycle. All possible origins need to be properly licensed, while only a subset is activated during S-phase. Origin licensing consists of the sequential loading of pre-RC proteins and starts with the binding of the origin recognition complex (ORC, Figure 1.1).

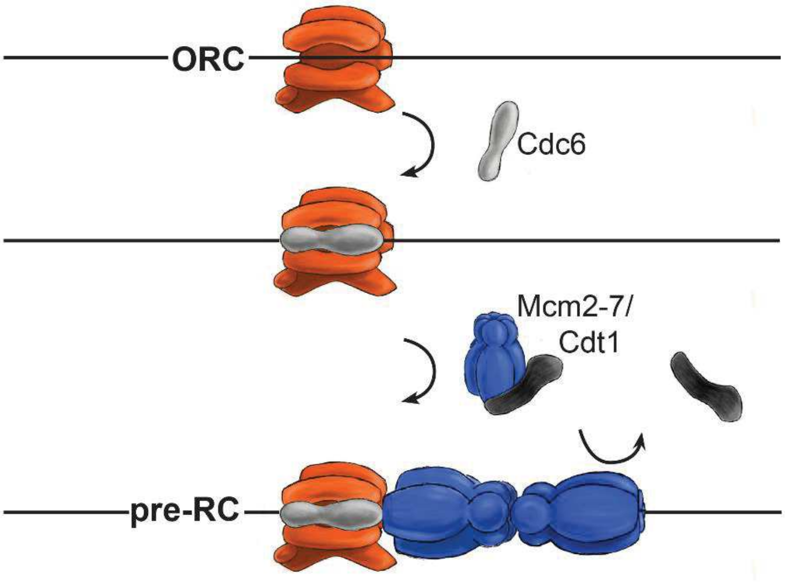


FIGURE 1.1: SCHEMATIC REPRESENTATION OF SEQUENTIAL PRE-RC ASSEMBLY. Licensing consists of the sequential loading of pre-RC components and is restricted to G1-phase. First ORC binds DNA as hexameric complex. Cdc6 recruitment traps DNA in the center of the circle and initiates Cdt1-dependent Mcm2-7 helicase loading. Cdt1 dissociates from Mcm2-7 after loading. Another Mcm2-7 hexamer is subsequently loaded in head-to-head conformation (Model template from Bleichert, Botchan, and Berger 2015).

ORC is a hexameric complex with ATPase activity and ATP-dependent DNA contact (Bleichert, Botchan, and Berger 2015). After ORC binding, the ATPase Cdc6 is recruited and DNA-bound ORC-Cdc6 initiates Cdt1 association. Cdt1 chaperones Mcm2-7 (for Minichromosome Maintenance) loading, the ring-structured core of the replicative helicase (Remus *et al.* 2009; Bell and Kaguni 2013). Cdc6 hydrolyses ATP only when bound to ORC and is required for proper Mcm2-7 loading. Mcm2-7 helicases are sequentially loaded as head-to-head double-hexamer (Remus *et al.* 2009). Interestingly, once Mcm2-7 double-hexamers are loaded, ORC, Cdc6, and Cdt1 are no longer required for replication initiation (Hua and Newport 1998; Yeeles *et al.* 2015). Pre-RC assembly is restricted to G1-phase when cyclin-dependent kinase (CDK) activity is low. Licensed origins are then competent for replication activation.

1.1.2 REPLICATION ACTIVATION IN S-PHASE: FROM PRE-REPLICATION COMPLEX TO PRE-INITIATION COMPLEX

With the onset of DBF4-dependent kinase (DDK) and CDK activity during the G1/S-phase transition (for kinase activities, see also Figure 1.2, chapter 1.1.4, p. 124), pre-RCs are converted into the pre-initiation complexes (pre-ICs). Pre-IC formation involves the binding of further proteins, such as Mcm10, Cdc45, Dbp11 and the GINS (Sld5, Psf1, Psf2, Psf3) complex (Gambus *et al.* 2006). During origin activation, several pre-IC proteins, including Mcm2-7, are phosphorylated by CDK and DDK to initiate replication (Francis *et al.* 2009). Cdc45-Mcm2-7-GINS (CMG) represent the active replicative helicase, which unwinds the DNA and DNA polymerase loads on single-stranded DNA (Ilves *et al.* 2010). Replication activation involves the dissociation of the Mcm2-7 double-hexamer into two active hexamers that originate the two replisomes consisting of about 150 proteins replicating DNA bidirectionally (Herrera *et al.* 2015).

Bidirectional DNA replication involves one strand to be replicated continuously (leading strand), while the other strand is synthesized in discontinuous fragments (Okazaki fragments), which are subsequently ligated (lagging strand). RNA primase *de novo* generates RNA primers, which are extended by DNA polymerases in 5' to 3' direction and removed during a maturation process (see also Figure 1.6, p. 22). DNA polymerases duplicate several hundreds of kb before replication forks collapse (Fragkos *et al.* 2015; Petryk *et al.* 2016). Replication termination happens e.g. when two replisomes converge.

1.1.3 TERMINATING DNA REPLICATION

Although initiation of DNA replication is well characterized, replication termination remains poorly understood. Replication termination mainly occurs when two opposite replication forks (from two adjacent origins of replication) collide. Also, termination needs to be tightly regulated to prevent premature termination without completing replication. The process involves the disassembly of converging replisomes and resolution of replication-induced DNA catenated intertwines by topoisomerases (Fachinetti *et al.* 2010). Recent work has shown that the polyubiquitination of the Mcm2-7 subunit Mcm7 at the end of S-phase leads to CMG disassembly (Moreno *et al.* 2014; Maric *et al.* 2014). However, this ubiquitination is necessary but not sufficient, suggesting additional factors being involved in replication termination (Lengronne and Pasero 2014).

1.1.4 CELL CYCLE REGULATION OF DNA REPLICATION: PREVENTING RE-REPLICATION AND INCOMPLETE REPLICATION

It is crucial for genetic stability, that the genome is *completely* replicated only *once* per cell cycle. The proteins that regulate cell cycle progression (amongst others cyclins, CDKs, and ubiquitin ligases) are also tightly linked with the control of DNA replication (DePamphilis *et al.* 2006). Origin licensing only happens in absence of DDK and CDKs (Figure 1.2). Existing pre-RCs are inactivated during S, G2 and M, preventing repeated licensing and activation (Blow and Dutta 2005; DePamphilis 2005). As already mentioned in chapter 1.1.2, DDK and CDK phosphorylate several members of pre-RC and pre-IC, which leads to replication activation, but also inhibits re-licensing of replication origins.

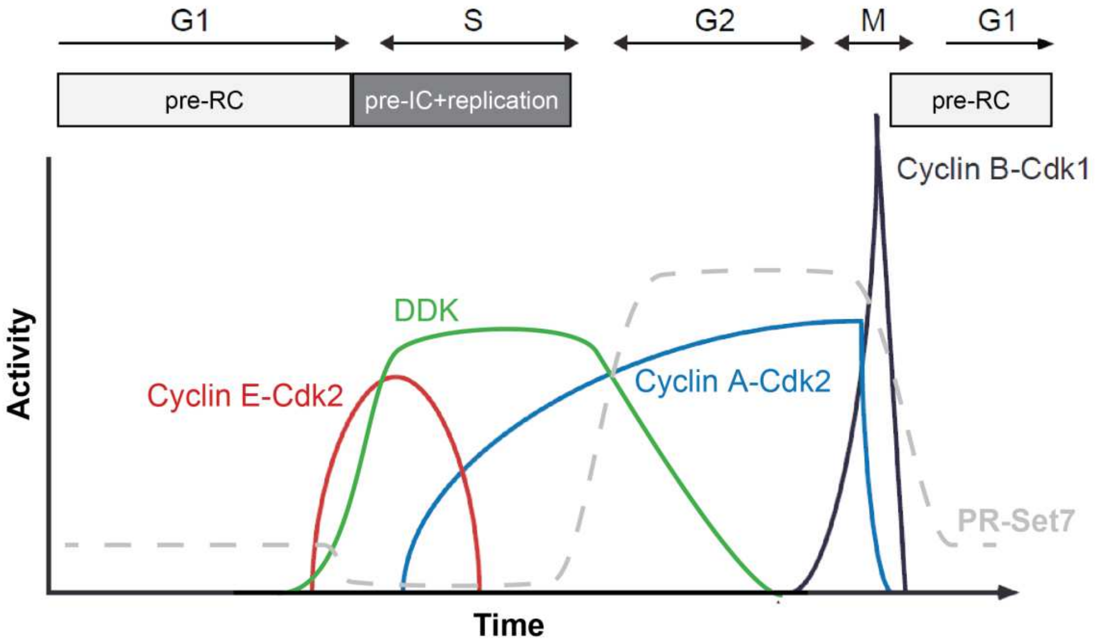


FIGURE 1.2: CELL CYCLE KINASE ACTIVITIES DETERMINE PRE-RC FORMATION AND ACTIVATION. Pre-RC assembly happens in absence of kinase activities. With onset of CDKs (red, blue) and DDK (green), replication is activated and re-licensing prevented (Adapted from PINES 1999). The dashed grey line represents expression H4K20 monomethyltransferase PR-Set7 (introduced in chapter 1.2.2, p. 18.)

In metazoans, Cyclin A-Cdk2 phosphorylates Cdt1 and Orc1 (Depamphilis *et al.* 2012). Phosphorylation of the Mcm2-7 chaperone Cdt1 in S-phase, leads to its ubiquitination, export to the cytoplasm, and subsequent degradation. This process ensures Cdt1 protein levels only being present in the nucleus in late mitosis/early G1. Additionally, Cdt1 activity is also repressed by Geminin, a specific Cdt1 inhibitor exclusively existing in metazoans. Geminin is active from S-phase on and degraded with the onset of mitosis. Interestingly, Cdt1 recruits Geminin to DNA and their binding stabilizes Cdt1, probably securing the availability of Cdt1 for the next G1-phase (Ballabeni *et al.* 2004).

In metazoan cells, ORC subunits are phosphorylated to prevent re-licensing during and after S phase (DePamphilis 2005; DePamphilis *et al.* 2006). Orc1 phosphorylation reduces binding affinities to chromatin is followed by ubiquitination and degradation (Méndez *et al.* 2002). Orc2 phosphorylation also dissociates ORC subunits 2-5 from DNA (Lee *et al.* 2012). Orc1 re-associates to chromatin during mitosis to G1 transition, followed by origin licensing (DePamphilis 2005).

Replication forks encountering obstacles (such as DNA secondary structures or lesions), or reduced deoxyribonucleotide triphosphate (dNTP) pools within the cell can lead to replication stress and induce replication fork stalling (Yekezare, Gómez-González, and Diffley 2013). For the cell to complete DNA replication, it is crucial to activate another licensed origin that has not been activated so far (dormant origin). Consequently, activation of additional origins decreases inter-origin distances and S-phase length remains constant (Blow, Ge, and Jackson 2011). Existence of dormant origins requires an excess of origin licensing during G1. Only a small proportion of licensed origins is actually activated during S-phase, while the vast majority remains dormant (Musiałek and Rybaczek 2015) and are evicted during passive replication (Kuipers *et al.* 2011). The choice of which origin to activate is thereby mainly stochastic.

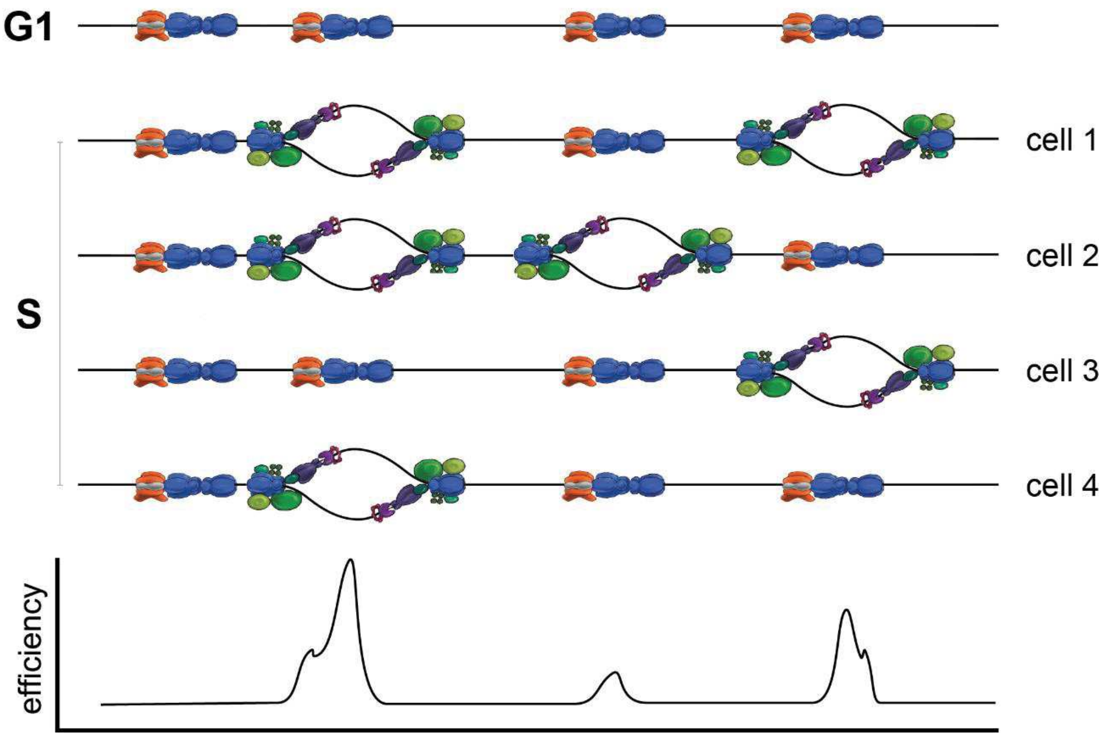


FIGURE 1.3: ORIGIN USAGE FREQUENCY DEFINES ORIGIN EFFICIENCY. Pre-RCs are assembled in G1. Only a subset is stochastically activated during S-phase, with each cell using a different cohort. The frequency of origin usage defines origin efficiency. Origin efficiency is represented schematically as result of SNS-seq (explained in chapter 1.4.2).

1.1.5 METAZOAN ORIGIN LICENSING AND ACTIVATION IS MAINLY STOCHASTIC

Only a subset of licensed origins is activated with the onset of S-phase and each cell uses a different cohort of replication origins. This observation led to the definition of origin efficiency as the frequency of a specific origin to be activated in a given cell during a given cell cycle (Méchali 2010) (Figure 1.3).

There are many features identified so far that influence origin licensing and efficiency (Figure 1.4). While AT-richness defines ARS elements in yeasts, there is a clear preference for CG-rich regions in mammals (Cayrou *et al.* 2015). Interestingly, not only CG-content, but also potentially resulting three-dimensional G-quadruplex structures (G4) impact on origin efficiency (Besnard *et al.* 2012; Valton *et al.* 2014; Langley *et al.* 2016). Thereby it remains controversial whether G4s actually positively or negatively influence DNA replication (Valton and Prioleau 2016). DNA accessibility, as well as specific histone modifications have been linked to active replication (introduced in detail in chapter 1.2, p. 17), as well as transcriptional activity and enhancer functions. Origins are often more concentrated and active in promoter regions, most probably due to open chromatin configurations, as direct interactions of transcription factors and replication factors have not been found so far (Fragkos *et al.* 2015). However, while all these different features contribute to origin licensing and activation on different levels, none of them is sufficient on its own and it is most likely a random combination of these different features that defines an origin (Méchali 2010).

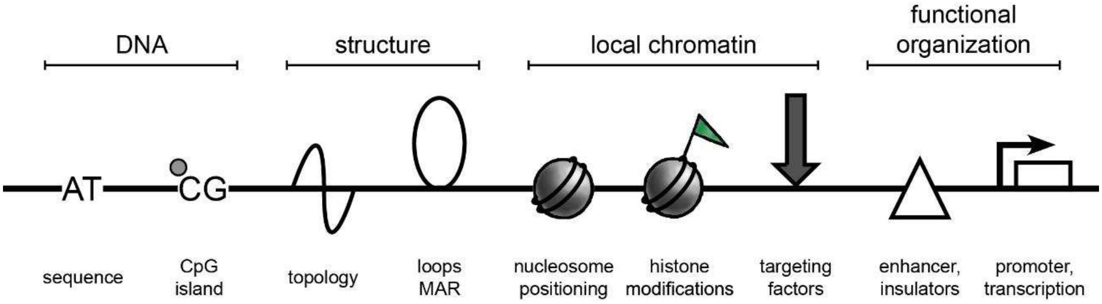


FIGURE 1.4: DIFFERENT FEATURES INFLUENCE REPLICATION ORIGINS. From DNA sequence to structure over chromatin to functional organization, all features have been linked to regulation of replication, while none of these features alone define an origin. DNA: AT-rich sequences might facilitate replication activation due to lower melting temperatures. Replication and GC-richness seems to be mainly linked on the structural level, as G4-structures are enriched at active initiation sites. Whether G4-structures facilitate pre-RC binding or replication initiation remains unclear so far. Structure: Bent DNA or loop formation might also facilitate pre-RC loading; MAR: matrix attachment regions. Local chromatin: nucleosome positioning/ nucleosome-free regions promote pre-RC binding, specific histone modification might directly interact with pre-RC components, targeting factors (EBNA1, HMGA1a) interact with ORC and direct replication. Functional organization: Active transcription/transcriptional regulation through enhancers/promoters have been described to influence DNA replication. Direct interactions between replication origin factors and transcription factors have not been reported. (Modified from Méchali 2010).

Recent work on the sequencing of Okazaki fragments led to the identification of broad (10-100 kb) initiation zones (Petryk *et al.* 2016, introduced in detail in chapter 1.4, p21). The authors claim that broad initiation zones represent *replication units*, within which replication preferentially initiates from multiple inefficient origins but only one single origin stochastically fires per cell. This observation also implies the necessity of broadly distributed licensed replication origins within such replication units.

It has been demonstrated, that the number of chromatin-bound Mcm2-7 helicases exceeds the number of active origins of ORC by a factor of 10 to 50 (Donovan *et al.* 1997; Powell *et al* .2015; Hyrien 2016). This can be achieved by either i) ORC loading several Mcm2-7 double-hexamers which spread from their binding site (Powell *et al.* 2015, *Drosophila*) or ii) ORC loading only one Mcm2-7 double-hexamer at a time, associating and dissociating quickly from DNA (Sonneville *et al.* 2012, *C. elegans*). There is evidence for either mechanism and further investigations will be required to conclusively resolve this question.

Finally, these recent findings evoked the model of the Mcm2-7 double-hexamer excess influencing organization of replication timing (Das and Rhind 2016; Hyrien 2016). Replication timing describes the time-point of origin activation during S-phase (early, middle, or late). Given that Mcm2-7 helicase activation can occur without the presence of ORC, higher densities of Mcm2-7 proteins could define early replicating regions, as the probability of early origin firing is simply higher than in late replicating domains containing less Mcm2-7 double-hexamers loaded on DNA. The organization of replication timing domains is described in chapter 1.3.

1.2 DNA REPLICATION REGULATION THROUGH CHROMATIN

Chromatin is a combination of DNA and proteins and ensures DNA compaction in the cell. The core unit of chromatin is the nucleosome – 147 bp DNA wrapped 1.7 times around a histone octamer consisting of two copies of H2A, H2B, H3 and H4. The most accessible chromatin level is the 10 nm fiber, often also referred to as "beads on a string" while higher order compacted structures constitute condensed chromatin (Soshnev, Josefowicz, and Allis 2016). Two main chromatin states exist: i) the compact and transcriptionally inactive state of chromatin is called heterochromatin, while ii) euchromatin represents accessible, transcriptionally active chromatin. Specific modifications of histone tails regulate chromatin states. Activating histone marks often involve acetylation, methylation (e.g. H3K4me1/2/3), and ubiquitination (Kouzarides 2007). Main repressing histone modifications associated to gene silencing are H3K9me3 and H3K27me3. Histone modifications are recognized by specific chromatin readers, which recruit other chromatin modifiers or remodelers, establishing chromatin as a very dynamic structure, being able to rapidly respond to environmental cues and requirements. However, not only histone modifications regulate chromatin structure, also nucleosome positioning impacts on chromatin and gene expression. Promoters and enhancers frequently exhibit accessible chromatin conformations, to allow binding of transcriptional regulators (Kouzarides 2007). Chromatin accessibility is not only a feature of transcriptional regulations, also pre-RC proteins preferentially bind accessible DNA regions (Méchali 2010; Miotto, Ji, and Struhl 2016), rendering chromatin conformation also important for regulation of DNA replication.

1.2.1 HISTONE MODIFICATIONS REGULATING DNA REPLICATION

Nucleosome depleted regions (NDRs) mark origins of replication in yeast (Field *et al.* 2008), Drosophila (Ding and MacAlpine 2011), Epstein-Barr-Virus (Papior *et al.* 2012) and humans (Miotto, Ji, and Struhl 2016). Nucleosome depletion might also be the link between DNA replication and G4 structures, as G4s exclude nucleosomes, thereby eventually favoring pre-RC formation (Fenouil *et al.* 2012; Valton and Prioleau 2016).

Transcriptionally active chromatin is marked by active histone modifications, typically acetylations of lysine residues of histone H3 and H4 (H3ac, H4ac), and H3K4me1/2/3. Genome-wide studies often correlate active chromatin modifications with DNA replication (Cadoret *et al.* 2008; Sequeira-Mendes *et al.* 2009; Valenzuela *et al.* 2011; Martin *et al.* 2011; Picard *et al.* 2014; Smith *et al.* 2016; Miotto, Ji, and Struhl 2016). However, most of these correlations result from the general association of replication with accessible chromatin. The only evidence of direct interactions between pre-RC components and chromatin regulators is the histone acetyltransferase HBO1 (histone acetyltransferase binding to Orc1, (Iizuka *et al.* 2009; Miotto and Struhl 2010). HBO1 interacts with Orc1 and Cdt1, and acetylates H4K5 and H4K12 in G1, which leads to chromatin decondensation and is necessary for Mcm2-7 loading.

In silent chromatin, origin licensing seems to be differently organized. Compacted chromatin has no accessible regions for pre-RCs to bind. Peptide-binding assays reported H3K9me3 and H3K27me3 peptides to interact with ORC components (Vermeulen *et al.* 2010). It has been shown very recently that pre-RC proteins directly interact with H3K9me3 demethylase Kdm4d, which presumably removes H3K9me3 to prepare favorable chromatin environment for origin firing (R. Wu *et al.* 2016). However, the question remains of how pre-RC assembly occurs in heterochromatin in the first place. One candidate is H4K20me3, which also localizes in silent chromatin and has been shown to directly interact with ORC (Vermeulen *et al.* 2010; Kuo *et al.* 2012; Beck *et al.* 2012).

1.2.2 HISTONE 4 LYSINE 20 METHYLATION AFFECTS DNA REPLICATION

Histone 4 lysine 20 methylation (H4K20me) is for several reasons the most promising histone methylation to be directly involved in regulation of DNA replication. First, the H4K20 monomethyltransferase PR-Set7 (also known as Set8, SetD8 or KMT5A) is cell cycle-dependently regulated, with low protein levels in G1 and complete absence in S-phase, while expression increases in G2 and peaks in mitosis (Figure 1.2, grey dashed line; S. Wu *et al.* 2010; S. Wu and Rice 2011). Second, both stabilization of PR-Set7 expression and depletion of PR-Set7 have severe effects on DNA replication and S-phase progression. Cell cycle regulation of PR-Set7 occurs mainly through the CRL4^{Cdt2} E3 ubiquitin ligase complex that uses PR-Set7 as direct substrate and targets it for proteasomal degradation during S-phase (Abbas *et al.* 2010; Centore *et al.* 2010; Oda *et al.* 2010). Expression of a non-degradable mutant (PR-Set7^{PIPmut}) leads to DNA re-replication, suggesting repeated origin licensing and activation (Tardat *et al.* 2010; Beck *et al.* 2012). Additional mutation of the SET-domain of PR-Set7, responsible for methylation activity, rescues this re-replication phenotype. Complete absence of PR-Set7 impairs cell cycle progression and is embryonic lethal at early stages of mouse development (Oda *et al.* 2010; S. Wu and Rice 2011; Beck *et al.* 2012). PR-Set7 is the only enzyme known to catalyze H4K20me1, whereas Suv4-20h1 and –h2 further convert H4K20me1 to H4K20me2 and –me3 (Schotta *et al.* 2004; Schotta *et al.* 2008; Brustel *et al.* 2011). Interestingly, loss of Suv4-20h1/h2 and H4K20me2/3 results in a less severe phenotype than PR-Set7 knock-out (Schotta *et al.* 2008), strengthening the importance for H4K20 methylation states during cell cycle regulation.

Artificial targeting of PR-Set7 to an integrated targeting site in the genome leads to induction of H4K20me1, conversion to H4K20me2/3 by endogenous Suv4-20h1/2 enzymes and to pre-RC assembly at the targeting site (Tardat *et al.* 2010). However, while Beck *et al.* claim necessity of conversion to H4K20me2/3 for efficient origin licensing (Beck *et al.* 2012), half of all detected active origins are found associated to H4K20me1 (Picard *et al.* 2014). Consequently, the exact role of PR-Set7/H4K20 methylation in origin licensing and/or activation remains to be uncovered.

1.3 DNA REPLICATION DEPENDS ON NUCLEAR CHROMATIN ORGANIZATION

DNA replication is organized on three different spatial levels. The first organizational level was described in detail in the chapter 1.1, consisting in pre-RC formation and single origin activation. The second level is composed in replication units (50-120 kb in size), which contain several origins from which only one is flexibly activated per cell. The third level consists of synchronously firing clusters of active replication units, the replication domains (400 kb – 1 Mb), which depend on nuclear compartmentalization (Fragkos *et al.* 2015).

Replication domains are mainly characterized by their replication timing program. Replication timing is established at one precise point during G1 (timing decision point), which coincides with chromatin anchorage after mitosis (Pope and Gilbert 2013). Concordantly, replication domains accord with topologically associating domains (TADs) (Pope *et al.* 2014). TADs were described by recent Hi-C (evolution of 3C: chromosome confirmation capture technique) experiments and are units of chromatin in spatial proximity, which are separated by distinct conserved boundaries. These boundaries are defined at the molecular level by CCCTC-binding factor (CTCF) and cohesins (Ciabrelli and Cavalli 2015). Interestingly, TADs are inherited to daughter cells after cell division. Furthermore, they are often conserved across cell types and species (Gonzalez-Sandoval and Gasser 2016) and are only reorganized during lineage specification in embryonic stem cells (Wilson *et al.* 2016). TADs contacting the nuclear lamina are also called LADs (lamin associated domains). LADs represent repressive chromatin with low transcriptional activities (Guelen *et al.* 2008). In general, LADs and also TADs with low gene densities and heterochromatin marks are replicated late in S-phase, while early replicating TADs are gene-rich, transcriptionally active with active histone modifications (Fragkos *et al.* 2015) (Figure 1.5). With the compartmentalization of the genome into replication domains, which are further subdivided into replication units, cells reduce the complexity of replication origin distribution. Also, replication fork stalling can be resolved at a local level and avoids involving the whole cellular organization (Rivera-Mulia and Gilbert 2016b).

In conclusion, replication timing is largely imposed by the nuclear compartmentalization of chromatin in TADs and their transcriptional activity. Thus, chromatin state within these TADs also defines origin activities. Boundaries of these domains are precisely defined by insulators (CTCF, cohesins) and are also hypersensitive to DNaseI, indicating a nucleosome- and protein-free chromatin environment.

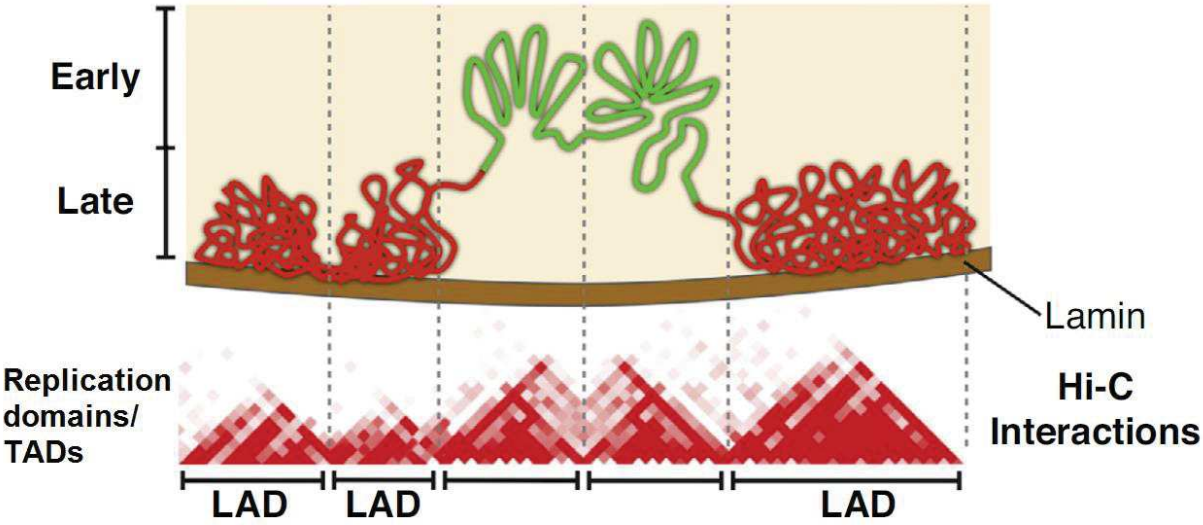


FIGURE 1.5: REPLICATION DOMAIN MODEL. Upper panel: Active chromatin TADs replicate early in S-phase and locate towards nuclear interior, while heterochromatin TADs replicate late during S-phase and locate towards the nuclear periphery (these lamin-associated TADs are also called LADs). Lower panel: Replication domains correspond to TADs. TADs are generally identified using Hi-C techniques and interactions between regions are plotted as Hi-C interaction heatmaps with dark-red regions strongly interacting and light-red regions only showing weak interaction (from Rivera-Mulia and Gilbert 2016a).

1.4 MAPPING OF DNA REPLICATION INITIATION EVENTS

Several different approaches have been developed to detect active replication initiation events in metazoans. Strategies vary from detection of single, defined replication origins to revelation of whole replication units.

1.4.1 SINGLE MOLECULE DNA COMBING

Single molecule DNA combing consists in pulse-labeling replicating DNA with the thymidine analogues CldU and IdU and detecting progress of replication forks by fluorescent microscopy. This technique requires uniform spreading of DNA molecules on a cover-slip and allows obtaining precise information about replication fork speed, fork asymmetry and inter-origin distances within the same DNA molecule during the same S-phase. Combining this technique with fluorescence *in situ* hybridization (FISH) allows to map precise loci of interest (Urban *et al.* 2015). However, technical limitations permit only detecting a defined set of single origins and efforts are currently made to establish high-throughput origin detection by molecular combing (De Carli *et al.* 2016).

1.4.2 BUBBLE-SEQUENCING AND SHORT NASCENT STRAND SEQUENCING

Alternative methods taking advantage of high-throughput sequencing are bubble trap or the purification of short nascent strands (SNS) followed by sequencing. Bubble trap uses the circular nature of replication bubbles to trap restriction fragments containing replication forks in gelling agarose. Purification and subsequent sequencing revealed 35000 to 40000 bubble-containing fragments per experiment (Mesner *et al.* 2013). However, although bubble trap was shown to contain few false positives (Mesner, Crawford, and Hamlin 2006), use of single restriction fragments does not allow to recover all possible genomic origins and resolution is limited (Urban *et al.* 2015).

For SNS-seq, short DNA fragments originating from leading strands of active replication forks are isolated and sequenced. Application of elevated sequencing depth revealed more than 200000 activated origins from bulk cells, representing different cohorts of potential origins being used per cell (Besnard *et al.* 2012). In short, this method relies on isolation of 500 - 2500 bp fragments by size fractionation on a sucrose gradient. Exploiting the RNA-primer used to initiate DNA replication, λ-exonuclease selectively digests all contaminating DNA fragments which are not protected by an RNA-primer from the 5′ end. The resulting DNA population represents nascent leading strands adjacent to replication origins (Figure 1.6). As λ-exonuclease digest remains debated in the field for possible biases (favoring detection of GC-rich/G4 DNA), nascent DNA can be alternatively labeled with the thymidine analogue BrdU and precipitated by immunoprecipitation (Fu *et al.* 2014). However, both methods have been shown to be highly concordant (Smith *et al.* 2016) and the detection of G4-structures at origins has been recently confirmed by an independent approach (Langley *et al.*

2016). Isolation of SNS followed by sequencing has been performed on several different organisms, including mouse (Cayrou *et al.* 2015) and human (Besnard *et al.* 2012; Picard *et al.* 2014; Cayrou *et al.* 2015; Smith *et al.* 2016). Thereby, origin efficiency can be deduced from the accumulation of sequencing reads per origin (schematically represented in Figure 1.3).

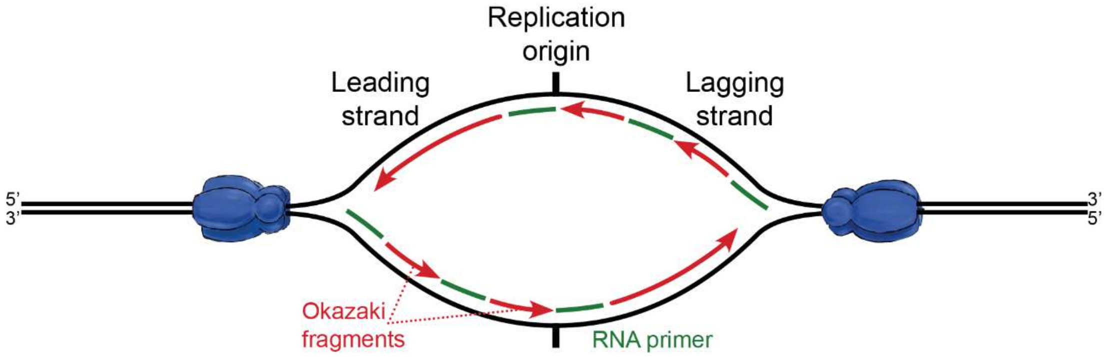


FIGURE 1.6: BIDERECTIONAL DNA REPLICATION REQUIRES LEADING AND LAGGING STRAND SYNTHESIS. RNA primer allow DNA polymerase to start synthesis. DNA polymerase only synthesizes in 5' → 3' direction, which leads to the continuously synthesized leading strand and the discontinuous Okazaki fragments.

1.4.3 OKAZAKI-FRAGMENT SEQUENCING

Isolation and sequencing of Okazaki fragments (OK-seq) not only represents an alternative method to detect replication origins, it also allows determining replication fork directionality (Petryk *et al.* 2016). Principal of OK-seq is pulse-labeling of Okazaki-fragments with the Thymidine analogue EdU and subsequent size fractionation of these fragments (< 200 bp). After biotinylation, adapter ligation, and precipitation, Okazaki fragments are sequenced and specific adapters allow to distinguish between Watson (leftward moving fork) and Crick (rightward moving fork) Okazaki fragments. Calculation of replication fork directionality (RFD = $\frac{(Crick-Watson)}{(Crick+Watson)}$) results in ascending (AS), descending (DS), and flat segments of different sizes and slopes (see also Figure 4.17 A). AS represent zones of preferential replication initiation, while DS are zones of preferential replication termination. The amplitude thereby reflects initiation efficiency. A broad initiation zone represents a zone of preferential replication initiation with multiple inefficient initiation sites but only one single origin firing per cell, corresponding to replication units (chapter 1.1.5, p. 15).

Interestingly, these different methods of replication origin detection show little concordance. SNS-seq studies already show poor overlap between each other, which might result from insufficient sequencing saturation (Urban *et al.* 2015). Also comparison of bubble-seq and SNS-seq only results in little agreement on the local level of single origins (Mesner *et al.* 2013). By contrast, initiation zones detected by OK-seq better align to bubbles than to SNS (Petryk *et al.* 2016). These little concordances between different approaches might originate from divergent replication events considered. SNS-seq focuses on narrow sites of defined active replication events, while bubble-seq and OK-seq both detect broader regions which might contain several origins, but only one origin firing per cell. Especially OK-seq reveals zones of preferential origin activation within initiation zones and preferential replication termination in termination zones. However, this does not limit origins to initiation zones, as inefficient origins might also be present (and detected by SNS-seq) in termination zones.

Information about pre-RC sites would help in this context, to bridge the discrepancies of the different detection methods. To date, two attempts have been made to perform ChIP against ORC components in human cells and will be introduced in the following.

1.4.4 ChIP-sequencing of pre-RC components

Little is known about genome-wide pre-RC positioning in humans. Major difficulty is thereby the low enrichment over background that hampers precise pre-RC detection (Schepers and Papior 2010). Due to the low sequence specificity and potentially high on-and-off rates, chromatin association of metazoan ORC is dispersed, especially in comparison to other nuclear factors (such as transcription factors). To date, only two genome-wide studies are available that target either Orc1 (Dellino *et al.* 2013) or Orc2 (Miotto, Ji, and Struhl 2016) in human cells. Dellino *et al.* overcame the high background problem by performing Orc1 ChIP from low-density chromatin. Thereby, they detected a prominent association of ORC to transcriptional start sites (TSS). However, chromatin selection prior to ChIP also introduced a bias towards Orc1 sites belonging to early replication origins (Dellino *et al.* 2013). Furthermore, 44% of all Orc1 sites overlap with replication origins identified by SNS-seq (Picard *et al.* 2014) and Orc1 was rather found at borders of replication initiation zones determined by OK-seq (Petryk *et al.* 2016). Miotto *et al.* performed Orc2 ChIP-seq experiments in unfractionated chromatin. Orc2 peaks showed moderate concordance with replication origins detected from SNS-seq (13% of all SNS sites located within 1 kb distance of Orc2 binding sites, 41% within 10 kb distance, Miotto, Ji, and Struhl 2016). Comparing Orc2 positions with several chromatin features, the authors concluded that ORC positioning majorly depends on chromatin accessibility.

Taken together, these results reveal that while ORC ChIP-seq already shows a certain level of concordance with active replication data, the overall picture remains incomplete. As ORC does not bind any consensus sequence in humans, but seems to solely depend on chromatin accessibility, ORC positions likely vary from cell to cell, resulting in a scattered ChIP-seq profile. This phenomenon might be even more pronounced for Mcm2-7, as there is evidence for multiple Mcm2-7 helicases at individual origins, which might even translocate from their original loading site (Das and Rhind 2016; Hyrien 2016). Consequently, Mcm2-7 ChIP-seq would result in a broad Mcm2-7 distribution, mostly lacking clear peaks. This is presumably the reason why human Mcm2-7 positions have not been assessed so far. However, as Mcm2-7 helicases can activate replication without spatial proximity to ORC, Mcm2-7 positions might be the missing link to conclusively connect replication initiation SNS-seq and ORC ChIP-seq results.

1.4.5 EPSTEIN-BARR VIRUS GENOME AS MODEL OF HUMAN DNA REPLICATION

Until now, genome-wide pre-RC ChIP experiments in humans have been unsuccessful. My laboratory used Epstein-Barr virus (EBV) as model system to study the relation between pre-RC formation and origin activation. EBV infects human B-cells and establishes a persistent latent infection. During latency, the EBV genome is maintained autonomously in proliferating cells and is thereby replicated in synchrony with the host cell genome by the cellular replication machinery (J. L. Yates and Guan 1991). For autonomous maintenance of the EBV episome, the *cis*-acting element *oriP* is required. *OriP* consists of two distinct elements: the family of repeats (FR-element) and the dyad symmetry (DS) element. Both elements contain binding arrays for the viral transactivator EBNA1 (Epstein-Barr virus nuclear antigen 1). By binding of EBNA1 to the FR-element, the EBV genome is tethered to the host chromatin during chromosome segregation (Marechal *et al.* 1999; Sears *et al.* 2003; Sears *et al.* 2004). EBNA1 binding to DS targets ORC to *oriP*, designating *oriP* as exceptional origin, since origin licensing depends on direct interaction between DNA, a targeting factor and ORC (Schepers *et al.* 2001; Ritzi *et al.* 2003). Thus, *oriP* represents a very strong origin licensing site, however, this is not necessarily accompanied by efficient origin activation (Papior *et al.* 2012). My laboratory took advantage of full chromatinization of the EBV genome (166 kbp in size), and high EBV copy-numbers in the Burkitt's lymphoma cell line Raji. Performing ChIP against the pre-RC components Orc2 and Mcm3, SNS isolation and micrococcal nuclease (MNase) digest in Raji cells, followed by hybridization against a designated microarray, allowed extensive analysis of the relation between origin licensing, activation and nucleosome positioning in a cell cycle resolved manner. They found 64 pre-RC sites, which also highly correlated with origin activation and S-phase-specific MNase sensitivities. Origin activation efficiencies were moderately influenced by AT-richness, however pre-RC sites themselves were independent of any specific primary motifs (Papior *et al.* 2012).

EBV employs cellular replication machinery to duplicate its genome and correlation of pre-RC and replication initiation sites provided insights in our understanding of replication origin organization in mammalian cells. It also showed that ChIP of human pre-RC components is technically feasible and sensibilized for need of careful controls. The next step will be to perform similar analysis on the human genome to answer the question of similar pre-RC organization genome-wide.

2. AIM OF THE THESIS

While the regulation of replication initiation in humans is already extensively studied and led to the identification of several contributing features, full understanding of the regulation of DNA replication can only be provided by combining replication initiation with the regulation of pre-RC positioning. Especially Mcm2-7 positioning – which has not been assessed so far – is expected to bridge the discrepancies between current ORC ChIP-seq and replication initiation data. In my laboratory, human pre-RC ChIP has been technically established using the EBV genome as model system. Intent of my thesis was to adopt this pre-RC ChIP technique for genome-wide ChIP-seq in Raji cells, by targeting the two major pre-RC subunits ORC and Mcm2-7. Containing the EBV genome as internal reference, Raji cells present the perfect control for ChIP-seq result quality. Once established, ChIP-seq will be also performed on embryonic stem (hES) cells, in collaboration with my co-supervisor Dr. Jean-Marc Lemaitre (Genome and Stem Cell Plasticity in Development and Ageing, IRMB, Montpellier, France). When compared to replication initiation, information about pre-RC positions and features that drive this positioning will contribute to the fundamental understanding of the relation between origin licensing and activation. Furthermore, pre-RC positions will be attributed to chromatin features, such as histone modifications. There is functional evidence for H4K20 methylation to be involved in DNA replication, however, we lack molecular understanding. Thus, ChIP-seq will also be performed for H4K20me1 and –me3, in order to directly conclude for possible implications between either H4K20 methylation and origin licensing. Possible candidates will be validated using an EBV-based plasmid system established in our laboratory, to functionally confirm the relation between H4K20 methylation and replication.

Consequently, I aim to combine genome-wide pre-RC positioning with the regulation of replication initiation. Simultaneously, I will evaluate the implication of H4K20 methylation as a promising histone modification candidate for regulation of DNA replication licensing and activation both by genome-wide correlation studies and by functional approaches.

3. MATERIAL AND METHODS

3.1 MEDIA, MATERIAL, DEVICES, CHEMICALS AND AGENTS

In the following, cell culture media and supplements (Table 3.1), chemicals and agents (Table 3.2), Enzymes (Table 3.3), Kits (Table 3.4), Material (Table 3.5), and devices (Table 3.6) used in this work are listed.

Table 3.1 lists all media, supplements, antibiotics, and other substances that were used for cell culture.

TABLE 3.1: CELL CULTURE MEDIA, SUPPLEMENTS, ANTIBIOTICS AND AGENTS.

Cell culture media and supplements	**Distributor**
BD Matrigel	BD Biosciences
CryoStor CS10	STEMCELL, Canada
Dimethylsulfoxide (DMSO)	Carl Roth GmbH, Germany
DMEM, high glucose, L-Glutamine	Gibco, Thermo Fisher, USA
Essential 8 Basal Medium + supplement	Gibco, Thermo Fisher, USA
Fetal calf serum (FCS)	Lot BS225160.5, Bio&SELL, Germany
G418	Carl Roth GmbH, Germany
L-Glutamin	Gibco, Thermo Fisher, USA
Lipofectamine 2000	Invitrogen, Germany
MEM non-essential amino acids	Gibco, Thermo Fisher, USA
PBS Dulbecco, pH 7.2	Biochrom AG, Berlin
Penicillin Streptomycin	Gibco, Thermo Fisher, USA
RPMI 1640	Gibco, Thermo Fisher, USA
Sodium pyruvate	Gibco, Thermo Fisher, USA
TripLE Select	Life technologies, USA
Trypsin-EDTA	Gibco, Thermo Fisher, USA
Versene solution	Gibco, Thermo Fisher, USA
Zeocine	Invitrogen, Germany

In the following Table 3.2, all chemicals and agents used during this work are specified.

TABLE 3.2: CHEMICALS AND AGENTS.

Chemicals and agents	**Distributor**
Ampicillin sodium salt	Carl Roth GmbH, Germany
ATX Ponceau S red staining solution	Fluka Analytical, Germany
Bovine serum albumin	Sigma-Aldrich, Germany
Bradford reagent, Bio-Rad protein assay	Bio-Rad, Germany
Chloroform	Merck-Eurolab GmbH, Germany
Deoxycholic acid (DOC)	Sigma-Aldrich, Germany
Dithiothreitol (DTT)	Sigma-Aldrich, Germany
EGTA, Titriplex®	Merck Millipore, Germany
Ethanol	Merck-Eurolab GmbH, Germany
Ethylenediaminetetraacetic acid (EDTA)	Carl Roth GmbH, Germany
Formaldehyde, MeOH-free	Thermo Scientific, USA
Glycerol	AppliChem GmbH, Germany
Glycine	Carl Roth GmbH, Germany
HEPES	Sigma-Aldrich, Germany
Isoamyl alcohol	Merck-Eurolab GmbH, Germany
Methanol	Merck-Eurolab GmbH, Germany
NP-40 (Igepal CA-630)	Sigma-Aldrich, Germany
Phenol	Carl Roth GmbH, Germany
Polyacrylamide	Carl Roth GmbH, Germany
Propidium iodide	Sigma-Aldrich, Germany
Salmon sperm	Invitrogen, Germany
Sodium chloride	Merck-Eurolab GmbH, Germany
Sodium dodecylsulfate (SDS)	Serva Electrophoresis GmbH, Germany
Sodium lauroyl sarcosinate (Sarkosyl)	Sigma-Aldrich, Germany
Tris	AppliChem GmbH, Germany
Triton-X-100	Sigma-Aldrich, Germany
Tween-20	AppliChem GmbH, Germany

Table 3.3 shows all enzymes employed during the study.

TABLE 3.3: ENZYMES.

Enzymes	**Distributor**
Benzonase	Sigma-Aldrich, Germany
DNase	Roche, Germany
DpnI	New England Biolabs, USA
Protease inhibitor complete	Roche, Germany
Proteinase K	Roche, Germany
RNase, DNase-free	Roche, Germany

Kits used during this work are listed in Table 3.4.

TABLE 3.4: KITS.

Kits	**Distributor**
BD Stemflow^{TM} Human and Mouse Pluripotent Stem Cell Analysis Kit	BD Biosciences, Germany
NTB buffer	Macherey-Nagel, Germany
NucleoSpin Extract II Kit	Macherey-Nagel, Germany
Qubit HS dsDNA	Invitrogen, Germany
SYBR Green I Master	Roche, Germany

In Table 3.5, all material needed for the experiments are specified.

TABLE 3.5: MATERIAL.

Material	**Distributor**
AFA Fiber & Cap tubes (12x12 mm)	Covaris Inc., UK
Amersham Hybond ECL	GE Healthcare, Germany
CEA Blue Sensitive X-ray films	Agfa Healthcare, Germany
Cell culture dishes and 6-well plates	Nunc GmbH, Germany
Cell strainer, 40 µm, 100 µm	Corning Inc., USA
CoolCell LX Freezing Container	Sigma-Aldrich, Germany
Cryotubes	Nunc GmbH, Germany
Nalgen Nunc Cryo 1°C freezing container	Nunc GmbH, Germany

Table 3.6 names the devices used during this work.

TABLE 3.6: DEVICES.

Devices	**Distributor**
Beckman JE-5.0 rotor with a large separation chamber	Beckman-Coulter, Germany
Cole-Parmer Masterflex pump	Cole-Parmer, USA
FACSCalibur™	BD Biosciences, Germany
semiDry blotting system	Hoefer Scientific Instruments, USA
Covaris S220	Covaris Inc., Germany
NanoDrop ND-1000 Spectrometer	ThermoScientific, USA
Qubit fluorometer	Invitrogen, Germany
Roche LightCycler 480 System	Roche, Germany
Optimax X-ray film processor	Rotec GmbH, Germany
Electroporation system Gene-Pulser II	Bio-Rad Laboratories, USA

3.2 BIOLOGICAL METHODS

3.2.1 CELL CULTURE

Cells with corresponding AGV-internal identification number, a short description and the respective media are listed in Table 3.7. More detailed cultivation information is given in the following.

RAJI CELLS

Raji cells (ATCC) were cultivated at 37°C and 5% CO_2 in RPMI 1640 (Gibco, Thermo Fisher, USA) supplemented with 8% FCS (Lot BS225160.5, Bio&SELL, Germany), 100 Units/ml Penicillin/ 100 µg/ml Streptomycin (Gibco, Thermo Fisher, USA), 1x MEM non-essential amino acids (Gibco, Thermo Fisher, USA), 2 mM L-Glutamin (Gibco, Thermo Fisher, USA), and 1 mM Sodium pyruvate (Gibco, Thermo Fisher, USA). Cells were routinely diluted to $2x10^5$ cells/ml and maximally grown to a density of $5x10^5$.

ADHERENT HEK293 CELLS

HEK293 EBNA1$^+$ cells were cultivated at 37°C and 5% CO_2 in DMEM (Gibco, Thermo Fisher, USA) supplemented with 8% FCS (Lot BS225160.5, Bio&SELL, Germany), 100 Units/ml Penicillin/ 100 µg/ml Streptomycin (Gibco, Thermo Fisher, USA), and 220 µg/ml G418 (Carl Roth GmbH, Germany). Cells were grown to 80% confluence and routinely split 1:4. Therefore, cells were washed with PBS, treated with 0.25% Trypsin-EDTA (Thermo Fisher, USA) for 2 min at 37°C, carefully resuspended in new medium and seeded on a new culture dish.

Generating stable HEK293 cell lines

Cells were seeded in a 6-well (Nunc GmbH, Germany) to a density of 2x105 and transfected with 3 µg linearized expression plasmid (see also Table 3.7) using Lipofectamine2000 according manufacturer's instructions (Invitrogen, Germany). Transfected cells from one 6-well were plated in medium with 20 µg/ml Zeocine (Invitrogen, Germany) on three 150 mm culture dishes (Nunc GmbH, Germany) the next day. After two to three weeks, single colonies were selected and expanded in 6-well plates. Expression of the protein of interest was verified by immunoblot. Positive clones were expanded and frozen.

hES cells H9

HES cells were cultivated in standard conditions (Ludwig *et al.* 2006) at 37°C, 5% CO_2. The hES cell line H9 was cultivated on BD Matrigel (Basement membrane matrix, BD Biosciences) covered dishes. From BD Matrigel stocks, 72 µl aliquots were routinely prepared in chilled 2 ml Eppendorf tubes and stored at -20°C. For BD Matrigel preparation, one aliquot was thawed, resuspended in 1.5 ml Essential 8 Basal Medium (Gibco, Thermo Fisher, USA) and mixed with 4.5 ml medium (6 ml total). Of this mixture, 1 ml was distributed per 35 mm dish and incubated for at least 30 min at 37°C. For larger plates, surfaces and volumes were upscaled accordingly.

For passaging, cells were washed twice with PBS, 1 ml of Versene solution (Gibco, Thermo Fisher, USA) was added per 35 mm dish. Before cells detached, Versene solution was removed, new medium added and cells were gently detached from the dish. Cells were routinely diluted 1:6 once per week. Medium was changed every day.

Cryopreservation

HEK293 and Raji cells were concentrated by centrifugation (200g, 7 min, room temperature) and the cell pellet was resuspended in FCS 10% DMSO to a concentration of ~ $5x10^7$ cells/ml (HEK293) and $5x10^6$ cells/ml (Raji cells). 1 ml cell suspension was aliquoted in 2 ml cryotubes (Nunc GmbH, Germany) and slowly cooled at a rate of -1°C/ min in a Nalgen Nunc Cryo 1°C freezing container (Nunc GmbH, Germany) at -80°C. After few days, cells were transferred in liquid nitrogen for long term storage. For thawing, cells were incubated at 37°C in water bath, washed with 30 ml fresh pre-warmed medium and plated accordingly.

HES cells were detached as previously described and concentrated by centrifugation (300g, 5 min, room temperature). 10^6 cells were resuspended in 1 ml CryoStor CS10 (STEMCELL, Canada), transferred to 2 ml cryotubes, placed in CoolCell LX Freezing Container (Sigma, Germany) at -80°C. After few days, cells were transferred in liquid nitrogen.

TABLE 3.7: ESSENTIAL CELL LINE INFORMATION.

Cell lines (AGV identification)	**Description**	**Medium**
Raji (#1577)	EBV-containing Burkitt's lymphoma cell line	RPMI 1640, 8% FCS, 1% Penicillin/ Streptomycin, 2 mM Glutamin, 1% MEM Non-essential amino acids, 1 mM Sodium Pyruvate
HEK293 EBNA1$^+$(#1803)	Human embryonic kidney cells, stably expressing EBNA1	DMEM, 8% FCS, 1% Penicillin/ Streptomycin, 220 µg/ml G418
HEK293 EBNA1$^+$ Gal4 (#2116)	Generated from cell line #1803 by integrating expression plasmid p5237	DMEM, 8% FCS, 1% Penicillin/ Streptomycin, 220 µg/ml G418, 20 µg/ml Zeocine
HEK293 EBNA1$^+$ Gal4-Suv4-20h1 (#2680)	Generated from cell line #1803 by integrating expression plasmid p5572	DMEM, 8% FCS, 1% Penicillin/ Streptomycin, 220 µg/ml G418, 20 µg/ml Zeocine
HEK293 EBNA1$^+$ Gal4-PR-Set7 (#2113)	Generated from cell line #1803 by integrating expression plasmid p5235	DMEM, 8% FCS, 1% Penicillin/ Streptomycin, 220 µg/ml G418, 20 µg/ml Zeocine
HEK293 EBNA1$^+$ Gal4-PR-Set7^{SETmut} (#2188)	Generated from cell line #1803 by integrating expression plasmid p5236	DMEM, 8% FCS, 1% Penicillin/ Streptomycin, 220 µg/ml G418, 20 µg/ml Zeocine
hES cells H9 (cultivated at the IRMB, Montpellier)	Human embryonic stem cell line, cultivated from passage 15 to 24	Essential 8 Basal Medium + Essential 8 Supplement

3.2.2 CELL CYCLE FRACTIONATION BY CENTRIFUGAL ELUTRIATION

For centrifugal elutriation, 5x10^9 exponentially growing Raji cells were harvested, washed with PBS and resuspended in 50 ml RPMI 1640/ 8% FCS/ 1mM EDTA/ 0.25 U/ml DNaseI (Roche, Germany). Concentrated cell suspension was passed through 40 µm cell strainer and injected in a Beckman JE-5.0 rotor with a large separation chamber turning at 1500 rpm and a flow rate of 30 ml/min controlled by a Cole-Parmer Masterflex pump.

While rotor speed was kept constant, 400 ml fractions were collected at increasing flow rates (40, 45, 50, 60, 80, 100ml/min). Individual fractions were quantified, 5x10^6 cells taken for propidium iodide stain and subsequent FACS analysis, while the remaining cells were subjected to cross-link (chapter 3.2.6, p. 35).

3.2.3 FLUORESCENCE ACTIVATED CELL SORTING (FACS)

CELL CYCLE DETERMINATION BY FACS

Cells from elutriation fractions (5x10^6 cells) were washed with PBS, resuspended in 3 ml PBS (4°C) and fixed by addition of 7 ml ice-cold 100% EtOH for at least 30 min at -20°C. After centrifugation at 400g for 5 min at 4°C, the pellet was washed with PBS 4°C and resuspended in 500 µl PBS 4°C. After 5 min of RNase (200 µg/ml final) treatment, propidium iodide (0.5 mg/ml in PBS) was added to a final concentration of 50 µg/ml and cells were subjected to FACSCalibur^{TM} (BD Biosciences, Germany) detection by the FL2 channel.

DETECTION OF PLURIPOTENCY MARKERS BY FACS

hES cells were stained and assessed for pluripotency markers with the BD Stemflow^{TM} Human and Mouse Pluripotent Stem Cell Analysis Kit (BD Biosciences) according to manufacturers' instructions.

3.2.4 IMMUNOBLOT

VERIFICATION OF STABLE CELL LINES

Cell extracts were prepared using NP-40 lysation. Cells were washed with ice-cold PBS and 400µl NP-40 extract buffer (4°C) were added to 15 cm culture dish. Cells were scraped off the plate and transferred to 1.5 ml Eppendorf tube. After 30 s vortex and 15 min on ice, cell lysate was centrifuged (16.100 g, 10 min, 4°C) and supernatant was kept for experiments or long term storage at -20°C.

Protein concentration was determined using Bradford reagent (BioRad, USA) and 50 µg protein extract with 1X Laemmli was loaded on the gel.

Proteins were separated on a 10% SDS-polyacrylamide gel and blotted on Amersham Hybond ECL (GE Healthcare, Germany) membrane using semiDry blotting system (Hoefer Scientific Innstruments, USA). Successful protein separation was verified by ponceau stain (ATX Ponceau S red staining solution; Fluka Analytical, Germany) and the membrane was blocked for at least 30 min in PBS, 1% Tween-20 (AppliChem GmbH, Germany), 5% milk. Incubation with primary anti-Gal4 (DBD) (sc-577, Santa Cruz Biotechnology) antibody (1:400 in PBS 1% Tween-20) was performed for 16h at 4°C, membrane was washed 3x 10 min with PBS 1% Tween-20 and incubated with anti-rabbit-HRP (Jackson ImmunoResearch Inc., USA) 1:10000 secondary antibody in PBS 1% Tween-20, 2.5% milk for 1 h at room temperature. After repeated washing steps, revelation was done using ECL on CEA Blue Sensitive X-ray films (Agfa Healthcare, Germany).

NP-40 extract buffer: 150 mM NaCl, 50 mM HEPES, 5 mM EDTA, 0.1% (v/v) NP-40, freshly add 1x protease inhibitor complete (EDTA-free, Roche, Germany)

5X Laemmli buffer: 250 mM Tris pH 6.8 (2M), 10% SDS, 500 mM DTT, 25% Glycerol, 0.5% Bromphenolblue

10% polyacrylamide running gel: 10% polyacrylamide (Carl Roth GmbH, Germany), 3.4 mM SDS, 375 mM Tris pH 8.8

Stacking gel: 4% polyacrylamide (Carl Roth GmbH, Germany), 3.4 mM SDS, 125 mM Tris pH 6.8

1X running buffer: 192 mM Glycine, 24 mM Tris pH 7.4, 3.4 mM SDS

Blotting buffer: 1X running buffer, 20% MeOH

ECL solution: 1 ml solution A, 3 µl solution B

> ***Solution A:*** 100mM Tris pH 8.8, 200 mM p-cumaric acid, 1.25 mM Luminol
>
> ***Solution B:*** 3% (v/v) H_2O_2

HISTONE IMMUNOBLOT

RIPA protein extraction was performed by trypsinizing HEK293 cells, washing with ice-cold PBS, adding two pellet volumes RIPA extract buffer (+ 1X complete protease inhibitor, Roche) and incubation for 20 minutes on ice. After 30 seconds vortex, 50 U Benzonase (Sigma, Germany) were added for 15 minutes at room temperature. Extract was centrifuged (16100 g, 15 min, 4°C), supernatant transferred to a new tube and subjected to immunoblot or stored at 20°C for long term storage.

Extract (50 µg) was loaded on a 15% layered with 12.5% polyacrylamide gel and treated as described above. Antibody dilutions were applied as listed in Table 3.8.

RIPA extract buffer: 50 mM Tris-HCl pH 7.9, 150 mM NaCl, 1% NP-40, 0.5% DOC, 0.1% SDS

15% polyacrylamide gel: 15% polyacrylamide (Carl Roth GmbH, Germany), 3.4 mM SDS, 375 mM Tris pH 8.8

12.5% polyacrylamide gel: 12.5% polyacrylamide (Carl Roth GmbH, Germany), 3.4 mM SDS, 375 mM Tris pH 8.8

TABLE 3.8: HISTONE ANTIBODIES AND RESPECTIVE DILUTIONS FOR IMMUNOBLOT.

Target protein	**Antibody**	**Dilution in PBS 1% Tween**
H4K20me1	Cell Signaling 9724S	1:1000
H4K20me2	Cell Signaling 9759S	1:1000
H4K20me3	Cell signaling 5737S	1:1000
H3K4me3	Abcam 8580	1:1000
H3K9me3	Active Motif 39161	1:1000
H3K27me2/3	Active Motif 39536 (use anti-mouse-HRP secondary antibody)	1:1000
H4	Abcam 31830 (use anti-mouse-HRP secondary antibody)	1:1000

3.2.5 PLASMID ABUNDANCE ASSAY

GFP-positive reporter plasmids (1µg, Table 3.9) were transfected into HEK293 EBNA1^{+} cell line stably expressing the respective GAL4-fusion protein using Lipofectamine2000 (Invitrogen, Germany) according to manufacturer's instructions. Transfections with comparable efficiencies were verified by visualizing GFP-positive cells. Six days post-transfection, cells were harvested according to the HIRT protocol (Hirt 1966). After washing with PBS, cells were equilibrated in 5ml TEN buffer. TEN was removed and cells were collected in 1,5 ml TEN buffer and an equal volume of 2X HIRT buffer. The lysate was then incubated at 4°C for 16h, in the presence of 1.25 M NaCl. After centrifugation at maximal speed, 4°C for 1h, DNA was purified by phenol-chloroform extraction (with Phenol/ Phenol-Chloroform/ Chloroform-Isoamyl alcohol steps). After precipitation, DNA was digested with 40 U *Dpn*I (NEB, USA) in presence of 200µg/ml RNase (Roche, Germany). Digested DNA (300 ng) was electroporated into Electromax DH10B competent cells (Invitrogen, Germany) and ampicillin-resistant colonies, representing the number of recovered plasmids, were counted. The FR-DS plasmid was always transfected in parallel and the number of resulting colonies was used for normalization.

The Suv4-20h1/2 inhibitor A-196 (kind gift of Structural Genomics Consortium (SGC)) was added to a final concentration of 5 µM immediately after transfection to the cell culture medium and the cells were kept under 5% O_2 during A-196 treatment. Medium was exchanged every second day.

TEN buffer: 10 mM Tris-HCl pH 7.5, 1 mM EDTA, 150 mM NaCl

2x HIRT buffer: 1.2% SDS, 20 mM Tris-HCl pH 7.5, 20 mM EDTA

TABLE 3.9: REPORTER PLASMIDS USED FOR PLASMID ABUNDANCE ASSAYS AND PLASMID CHIPS.

Plasmid (AGV identification)	Plasmid name
p3230	FR-DS
p5233	FR-UAS-DS
p5234	FR-UAS
p3244.2	FR-ori^{RDH}
p5588	FR-UAS-ori^{RDH}

3.2.6 CROSS-LINK

RAJI CELLS (STANDARD PROTOCOL)

Cells were washed twice with PBS, resuspend in PBS to a concentration of 2x10^{7} cells/ml and passed through 100 µm cell strainer (Corning Inc., USA). An equal volume of PBS 2% methanol-free formaldehyde (Thermo Scientific, USA) was added and cells were fixed for 5 minutes on a roller at room temperature. The cross-linking reaction was then quenched with glycine (125 mM final concentration) and incubated for another minute on the roller. After washing once with PBS and once with PBS 0.5% NP-40, cells were resuspended in PBS containing 10% glycerol, pelleted and snap frozen in liquid nitrogen.

hES CELLS

HES cells are more fragile and need to be resuspended with care. Cells were covered with 10 ml TripLE Select (Life technologies, USA) on a 15 cm dish and incubated until cells detach. Cells were dissociated by carefully resuspending once, transferred to centrifugation tube and washed once with PBS. Cell concentration was adjusted to 2x10^{7} cells/ml in PBS, passed through 100 µm cell strainer and cross-link was performed according to the standard protocol.

HEK293

Transfected HEK293 EBNA1^{+} cell lines were trypsinized, washed twice in PBS and standard cross-link protocol was performed.

3.2.7 CHROMATIN IMMUNOPRECIPITATION (CHIP)

SONICATION

Cross-linked cell pellets were thawed on ice for 15 min and resuspended for lysis in LB3+ buffer to a final concentration of 2x10^{7} cells/ml. Sonication was performed in AFA Fiber & Cap tubes (12x12 mm, Covaris, Great Britain) at an average temperature of 5°C according the settings established for each cell line (Table 3.10) using the Covaris S220 (Covaris Inc., Great Britain).

TABLE 3.10: SONICATION SETTINGS ESTABLISHED FOR EACH CELL LINE.

Cell line	Sonication settings
Raji	100W, 150 cycles/burst, 10% duty cycle, 20 min (S-phase: 17 min)
hES	100W, 150 cycles/burst, 10% duty cycle, 14 min
HEK293 EBNA1$^+$ Gal4-fusion	150W, 200 cycles/burst, 20% duty cycle, 20 min

IMMUNOPRECIPITATION

After sonication, sheared chromatin was pre-cleared with 50 µl protein A beads (protein A Sepharose 4 Fast Flow, GE Healthcare, Germany; washed 3x in PBS, 50% bead slurry prepared) per 500 µg chromatin for 2h. Chromatin concentration was measured by NanoDrop ND-1000 Spectrometer (ThermoScientific, USA). An appropriate amount of chromatin was incubated with the respective antibody (for plasmid ChIPs see Table 3.11, for ChIP-seq see Table 3.12) for 16h at 4°C. BSA-blocked protein A beads (incubated for 2h on roller in blocking solution at 4°C; protein G beads (protein G Sepharose 4 Fast Flow, GE Healthcare, Germany) were used for antibodies raised in mouse) were then added (50 µl/ 500 µg chromatin) and incubated for at least 4h on orbital shaker at 4°C. Sequential washing steps with RIPA-150 mM NaCl, RIPA-300 mM NaCl, LiCl buffer and finally twice in TE (pH 8.0) buffer were performed. Immunoprecipitated chromatin fragments were eluted from the beads by shaking twice at 1200 rpm for 10 min at 65°C with 100µl of TE and 1% SDS. The elution was treated with 80 µg RNAse A (Roche, Germany) for 2h at 37°C and with 8 µg proteinase K (Roche, Germany) at 65°C for 16h. DNA was purified using the NucleoSpin Extract II Kit (and NTB binding buffer for SDS containing samples) according to manufacturer's instructions. Quantitative PCR analysis was performed as described in 3.2.8, p. 38 and quantitative PCR values were represented as fold enrichment relative to isotype IgG control. Chromatin sizes were verified by loading 1-2 µg chromatin on an 1.5% agarose gel. Samples intended for ChIP-seq were quantified using Qubit HS dsDNA (Invitrogen, Germany).

LB3+ buffer: 25 mM HEPES (pH 7.5), 140 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.5% Sarcosyl, 0.1% DOC, 0.5% Triton-X-100, 1X protease inhibitor complete (Roche, Germany)

Blocking solution: 0.5 mg/ml BSA, 30 µg/ml salmon sperm (Invitrogen, Germany), 1X protease inhibitor complete, 0.1% Triton-X-100 in LB3(-) buffer (without detergents)

RIPA-150 mM NaCl: 150 mM NaCl, 0.1% SDS, 0.5% DOC, 1% NP-40, 50 mM Tris (pH 8.0), 1 mM EDTA

RIPA-300 mM NaCl: 300 mM NaCl, 0.1% SDS, 0.5% DOC, 1% NP-40, 50 mM Tris (pH 8.0), 1 mM EDTA

LiCl: 250 mM LiCl, 0.1% SDS, 0.5% DOC, 1% NP-40, 50 mM Tris (pH 8.0), 1 mM EDTA

1X TE: Tris-EDTA pH 8.0 (Carl Roth GmbH, Germany)

TABLE 3.11: ANTIBODIES FOR PLASMID CHIPS IN HEK293 EBNA1^{+} GAL4-FUSION CELL LINES.

ChIP target	**Antibody**	**Amount**	**Chromatin**
Gal4	anti-Gal4 (Santa Cruz Biotechnologies, sc-577)	2.5 µg	250 µg
H4K20me1	anti-H4K20me1 (Diagenode, MAb-147-100)	2.5 µg	
H4K20me3	anti-H4K20me3 (Diagenode, pAB-057-050)	2.5 µg	
Mcm3	anti-Mcm3 SA8413	15 µl	
IgG	Rabbit IgG (Sigma, R2004)	2.5 µg	

TABLE 3.12: ANTIBODIES FOR CHIP-SEQUENCING.

ChIP target	**Antibody**	**Amount**	**Chromatin**
Orc2	anti-Orc2 SA93	15 µl	500 µg
Orc3	anti-Orc3 SA7976	15 µl	500 µg
Mcm3	anti-Mcm3 SA8413	15 µl	500 µg
Mcm7	anti-Mcm7 SA8496	15 µl	500 µg
H4K20me1	anti-H4K20me1 (Diagenode, MAb-147-100)	2.5 µg	250 µg
H4K20me3	anti-H4K20me3 (Diagenode, pAB-057-050)	2.5 µg	250 µg
H4	anti-H4 (Millipore, 05-838, clone 62-141-13	5 µl	250 µg
IgG	Rabbit IgG (Sigma, R2004)	10 µg/ 2.5 µg	500 µg/ 250 µg

3.2.8 QUANTITATIVE PCR (QPCR)

Quantitative PCR was performed Roche LightCycler 480 System and the SYBR Green I Master (Roche). 2 µl of ChIP elution were mixed with [5 µl 2xSYBR, 2.5 µl H2O, 0.5 µl 5 µM primer mix]. Amplification was performed using the Roche SYBR standard program Table 3.13. QPCR primers are listed in Table 3.14.

TABLE 3.13: ROCHE SYBR QPCR STANDARD PROGRAM.

	Temperature [°C]	Duration [s]	Cycles	Detection
Pre-incubation	95	300	1	
Amplification	95	1	45	
	60	10		
	72	10		
	75	3		single
Melting curve	97	1	1	
	67	10, then heat to 97°C		
	97			continuous
Cooling	37	15		

TABLE 3.14: QPCR PRIMERS.

Primer (Schepers group internal numeration)	**Sequence [5' → 3']**
DS_for (575)	AGTTCACTGCCCGCTCCT
DS_rev (576)	CAGGATTCCACGAGGGTAGT
FR_for (276)	CGTGCTCTCAGCGACCTCG
FR_rev (277)	TCAAACCACTTGCCCACAAAAC
UAS_for (270)	TTACAGTCCAAAACCGCAGGG
UAS_rev (271)	TTGTCGCTCCGTAGACGAAGC
ori^{RDH}_for (414)	CTGTCTTGGTCCCTGCC
ori^{RDH}_rev (415)	TGCCTTCTCCTTCTCATCC
Mcm4+_for (776)	CTGAAAGAAGCGGCACTGTC
Mcm4+_rev (777)	CTCACCAATCACAGCGGC
Mcm4-_for (778)	CCACCCAGGCATTGCTAAAG
Mcm4-_rev (779)	CCCCTCTATTTGCCGTTCCT
BANF1+_for (772)	CCTCCCTTGTCCGTCTTCTA
BANF1+_rev (773)	GACCCGGAACTCAAGGACTTA
H4K20me1+_GAPDH_for (629)	ATGCCTTCTTGCCTCTTGTC
H4K20me1+_GAPDH_rev (630)	AGTTAAAAGCAGCCCTGGTG
H4K20me3+_ZNF180_for (621)	TCTGAGCAGGGTTGCAAGTAC
H4K20me3+_ZNF180_rev (622)	AAGGAAATGATGCCCAGCTG

3.2.9 SEQUENCING

The *Genomics unit of LAFUGA* (headed by Helmut Blum, LMU Munich) performed ChIP sample library preparations from > 4 ng of ChIP-DNA using Accel-NGS® 1S Plus DNA Library Kit for Illumina (Swift Biosciences). 50 bp single-end sequencing was done with the Illumina HiSEQ 1500 sequencer to a sequencing depth of ~ 70 million reads (for pre-RC) and 35 million reads (for histone modifications).

3.3 BIOINFORMATICS

Programs used for bioinformatics analyses are listed in tTable 3.15.

TABLE 3.15: PROGRAMS USED DURING BIOINFORMATIC ANALYSIS. Program sources are indicated. `[Typo] marks terminal commands`.

Program	**Source**
Bowtie1 (1.1.1)	https://sourceforge.net/projects/bowtie-bio/files/bowtie/1.1.1/
R (3.2.3)	https://cran.rstudio.com/
R package GenomicRanges (1.22.4) *R package IRanges (2.4.8)* *R package S4Vectors (0.8.11)* *R package BiocGenerics (0.16.1)* *R package RColorBrewer (1.1-2)* *R package gplots (3.0.1)*	In R: `install.packages("`*`package`*`")` `library(`*`package`*`)`
MACS2 (8.1.2)	`pip install macs2`
HOMER (v4.8)	http://homer.salk.edu/homer/
T-PIC	Download scripts from http://www.math.miami.edu/~vhower/tpic.html
Bedtools (2.25.0)	https://github.com/arq5x/bedtools2/releases
CEAS (1.0.2)	http://liulab.dfci.harvard.edu/CEAS/download.html
Venn Diagram Plotter	https://omics.pnl.gov/software/venn-diagram-plotter

3.3.1 BOWTIE MAPPING OF SEQUENCING READS AGAINST THE GENOME

Fastq-files from sequencing were mapped against the human genome (hg19, GRCh37, version 2009), extended for the EBV genome (NC007605) using the bowtie command:

```
bowtie -m 1 index file.fastq
```

3.3.2 GENERATION OF PILEUP PROFILES

Pileup profiles were generated in R by extending 50 bp reads by 150 bp and calculating the number of reads per base. The coverage was either saved as .wig files for visualization in IGB or as .rda for coverage analyses.

3.3.3 MACS2, HOMER, AND T-PIC PEAK-CALLING

PEAK-CALLING ON THE EBV GENOME

Peak-calling on the EBV genome was performed using the following commands. Commands in italics are variable and adapted for each file.

- MACS2:

```
macs2 callpeak -t ChIP.bam -c input.bam -f BAM -g 1.75e5 -n
output_name -B -q 0.01 –nomodel
```

- HOMER:

1. create tag directory:

```
makeTagDirectory tags_name1 input.bam –single
makeTagDirectory tags_name2 ChIP.bam –single
```

2. perform peak calling:

```
findPeaks tags_name2 -i tags_name1 -style factor -size 200 -
fragLength 200 -inputFragLength 200 -C 0 > output_name.txt
```

3. convert .txt to .bed

```
pos2bed.pl output_name.txt > output_name.bed
```

- T-PIC:

T-PIC uses scripts running in perl and R (download scripts from http://www.math.miami.edu/~vhower/tpic.html). The scripts need to be placed the same folder as ChIP and input .bed files. Scripts have to be adapted for each filename and each genome (create_coverage.pl: `my $bed_filename = "ChIP.bed";` zeta.pl: `my $bed_filename = "ChIP.bed", my $input_filename = "input.bed"`). For T-PIC peak-calling on the EBV genome, both scripts also need to be adapted for detecting the EBV "chromosome" (`my @index_set = ("EBV")`).

PEAK-CALLING ON THE HUMAN GENOME

The HOMER peak-calling command for the human genome does not differ from EBV. For histone ChIPs, the `– style` parameter was changed to `“histone”`.

For T-PIC peak calling, the scripts run with the human and filenames have to be modified as described previously.

MACS2 peak-calling for hES cells samples was performed with the following command:

```
macs2 callpeak -t ChIP.bed -c input.bed -g hs --broad -B -f BEDPE
-n output_name
```

3.3.4 DEFINITION OF ORC, MCM2-7, AND PRE-RC

Complexes were calculated by first merging overlapping peaks, counting the number of overlapping events and retaining only the positions display sufficient overlaps:

3.3.5 PEAK COMPARISONS: JACCARD INDEX AND OVERLAPS

JACCARD INDEX

Jaccard index was calculated using the `“bedtools jaccard”` function applied for every possible ChIP combination. Heatmap representation was calculated in R.

OVERLAPS

Overlaps of peak/ complex positions were calculated using the `“bedtools intersect”` function. Thereby, overlaps of the query files were calculated in both directions, merged, and the unique positions of every file were also retained. The results were used for Venn Diagram generation using VennDiagramPlotter.

3.3.6 CALCULATING GENOMIC DISTRIBUTIONS WITH CEAS

The genomic distribution of peaks or complexes were calculated using the command

```
ceas --name=output_name -g hg19.refGene -b ChIP/COMPLEX.bed
```

Thereby, hg19.refGene was downloaded from http://liulab.dfci.harvard.edu/CEAS/download.html. The results of interest were extracted and used for differential representation in R.

4. RESULTS AND DISCUSSION

To date, little is known about the mechanisms that define pre-RC formation and activation in mammals. Because PR-Set7 and consequently H4K20 methylation clearly impact on DNA replication, I decided to receive a first impression of the relationship between replication licensing and H4K20 methylation by ChIP-seq in human cells. I targeted both H4K20me1 and H4K20me3, as well as four different members of the pre-RC complex, targeting the two major subunits ORC and Mcm2-7: Orc2, Orc3, Mcm3, and Mcm7. In the following chapters, I will refer to the setup of ChIP experiments in both human lymphoblastoid Raji cells and hES cells H9. Further, I will detail the analysis of sequencing results and resulting conclusions in Raji cells, and a functional approach to functionally validate the relation between H4K20 methylation and DNA replication.

4.1 SETTING UP CHIP-SEQ EXPERIMENTS IN HUMAN CELLS

ChIP is a sensitive technique to study protein-DNA interactions. To fix the current chromatin state of the cells, DNA and proteins within a distance of ~2-3 Å are covalently linked by formaldehyde. After chromatin fragmentation, specific antibodies are used to precipitate the target proteins together with the associated DNA fragments. These DNA fragments are then isolated and quantified either by qPCR at defined loci or by sequencing to detect all possible protein binding sites.

4.1.1 CONSIDERATIONS TO OPTIMIZE PRE-RC CHIP-SEQ EXPERIMENTS

To date, there are only two ChIP-seq studies of pre-RC proteins in mammalian cells, both targeting the pre-RC subcomponent ORC (Dellino *et al.* 2013; Miotto, Ji, and Struhl 2016). Despite much effort, members of the Mcm2-7 complex have not been successfully targeted so far. One of the major difficulties is the low enrichment over background, challenging identification of clear pre-RC binding sites. In contrast to transcription factor and histone ChIPs, which are rather easy to perform and very robust to methodological differences, I considered several optimization steps for pre-RC ChIPs (Figure 4.1):

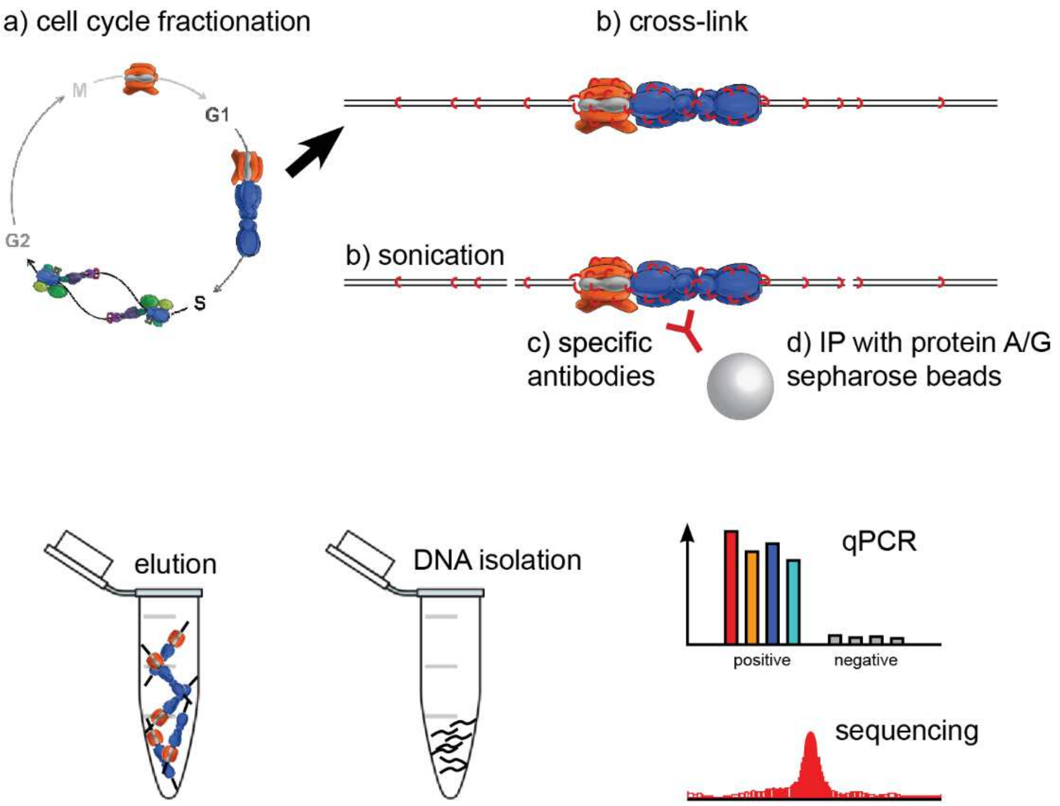


FIGURE 4.1: SCHEMATIC REPRESENTATION OF CRUCIAL CHIP STEPS. A) CELL CYCLE FRACTIONATION BY CENTRIFUGAL ELUTRIATION. B) CROSS-LINK AND SONICATION. C) ANTIBODY INCUBATIONS. D) IMMUNOPRECIPITATION AND WASHING CONDITIONS. After ChIP, DNA fragments were eluted from the beads, the DNA was isolated and quality control was performed by qPCR prior to sequencing.

a) Performing ChIPs on a G1 cell population
b) Optimizing cross-linking time and mild sonication to preserve complexes
c) Usage of several antibody targets within the same complex
d) Increasing antibody specificity (buffer and washing conditions)

Performing ChIP exclusively in G1 cell cycle stage - the moment where pre-RC is present on chromatin - is one possibility to increase sensitivity. Centrifugal elutriation is a convenient method to do so, as an asynchronous cell population is fractionized into the different cell cycle stages by simple size selection. This is easily feasible in Raji cells, however impossible in hES cells. I decided against chemical synchronization of hES cells and performed ChIPs on asynchronous cell populations instead.

ChIP procedure was optimized experimentally for increased stringency and sensitivity. First, cross-linking needs to be strong enough to efficiently conserve protein-DNA interactions, but should not impair proper sonication or mask any epitopes for antibody recognition. For this reason, I chose a very mild cross-link of 5 min with 1% FA at room temperature. Second, sonication itself needs to be as mild as possible, to not destroy any protein-DNA interactions, but should result in chromatin fragments small enough for high-throughput sequencing (200-400bp). The focused-ultrasonicator S220 from Covaris was used, as energy focusing directly on the sample allows to considerably reduce the applied power. This ensures maximum conservation of protein-DNA interactions. Indeed, usage of Covaris S220 in comparison to standard sonication methods (Bioruptor or tip sonifier) clearly increased protein enrichments compared to mock ChIPs at known positive loci (data not shown). Prior to the actual immunoprecipitation, chromatin was treated first with mock IgG antibody, then with protein A sepharose beads. This step removes background binding of chromatin fragments that stick to IgG and/or protein A sepharose beads.

To ensure unbiased detection of pre-RC sites, I simultaneously targeted four different pre-RC components: Orc2, Orc3 as members of ORC and Mcm3, Mcm7 as members of the Mcm2-7 complex. Targeting two different subunits decreases possible antibody biases and increases the significance of the results. The ChIP-grade of rabbit pre-immune serum antibodies targeting Orc2, Orc3, Mcm3, and Mcm7 have been already validated in repeatedly (Papior *et al.* 2012; Ghosh *et al.* 2006; Ritzi *et al.* 2003; Schepers *et al.* 2001). Furthermore, ChIPs were performed in three independent experiments to account for experimental and biological variation. During data analysis, usage of rigorous parameters for pre-RC definition from all these replicates increased confidence in the results.

Sonication and immunoprecipitation were performed in detergent containing buffer. I decided for a combination of Deoxycholate, Sarkosyl, Triton-X-100, and NP-40. After the precipitation, I performed sequential stringent washing steps with RIPA buffer containing SDS, NP-40, DOC, while increasing salt concentrations decrease low affinity interactions.

After protein-DNA complex elution from beads, samples were treated with RNase, covalent cross-link was reversed by heating to 65°C and proteins were digested using Proteinase K. The resulting DNA fragments were purified and inserted in control qPCR reactions.

4.1.2 PERFORMING CHIP-SEQ EXPERIMENTS IN RAJI CELLS

After careful setup of the ChIP conditions, I performed the ChIPs intended for sequencing. To start with, I fractionized asynchronous cell populations to obtain G1- and S/G2-phase enriched cells. The S/G2 population thereby served as control for pre-RC ChIP-seq, as activated and passively replicated pre-RCs disassemble after replication.

EXPERIMENTAL SETUP: CELL CYCLE FRACTIONATION

Counterflow centrifugal elutriation is a convenient method for separation of mixed cell populations according to size and mass and regularly used in my laboratory (Ritzi *et al.* 2003). Elutriation has the advantage to not disturb the cellular metabolism. The principle is based on centrifugation of a logarithmically growing asynchronous cell population, with largest cells (late cell cycle stages) being sedimented, while a counterflow of medium extracts smaller cells (early cell cycle stages) from this population. Regulation of the speed of this counterflow thus determines sizes of isolated cells. I routinely used counterflow rates 40, 45, 50, 60, 80, 100ml/min to isolate all different cell stages. The respective cell cycle stages were validated by propidium iodide (PI) staining of the DNA content and subsequent FACS analysis (Figure 4.2). I performed three independent elutriation experiments as basis for three ChIP replicates. I chose fractions G1 (counterflow rate 40 ml/min) - pre-RC formation is supposedly completed - and S/G2 cells (80 ml/min), directly after replication. The sonication time had to be slightly reduced for S/G2-phase cells, as the chromatin of S/G2 cells is fragmented more easily. An example for efficient sonication is depicted in Figure 4.3 A.

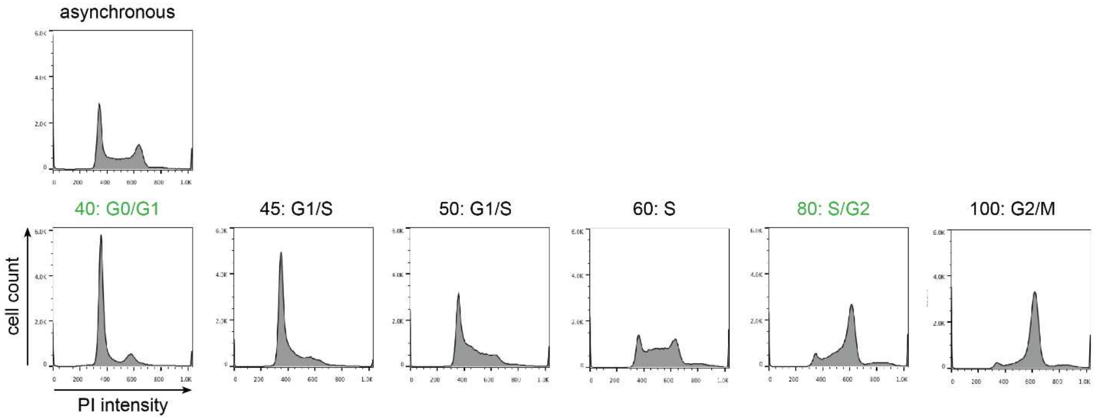


FIGURE 4.2: EXAMPLE FACS PROFILE OF RAJI CELLS FRACTIONNATED BY CENTRIFUGIAL ELUTRIATION REVEALED DISTINCT CELL CYCLE STAGES. Propidium iodide (PI) is an intercalating agent and was used to determine DNA content in the cells. The first peak represents the G1 cell population with 2n content, the second peak represents G2 population with 4n. The different counterflow rates led to a separation into cell populations of different cell cycle stages (indicated on top of each profile). The cell populations marked in green were taken for ChIP-seq experiments.

EXPERIMENTAL SETUP: CHIP VALIDATION

Like almost all EBV-infected cell lines, the Raji cell line maintains the EBV genome as an independent genetic entity, which is autonomously replicated during latency by the host cellular replication machinery (John L. Yates 1996). The latent Epstein-Barr virus origin *oriP* can serve as a positive control for pre-RC ChIP efficiencies. The viral latent protein EBNA1 specifically targets ORC to the dyad symmetry (DS) element of *oriP*, independent of the cell cycle stage. Consequently, ChIPs showed an enrichment of Orc2 and Orc3 at DS in both G1 and S/G2, while Mcm3 and Mcm7 binding is cell cycle dependent (Ritzi *et al.* 2003) and was decreased in S/G2 (Figure 4.3 B).

H4K20 methylation ChIPs were validated at previously identified genomic positive loci for either H4K20me1 or –me3 (primer sequences obtained from Stanimir Dulev, OICR, Canada). ChIP against canonical histone H4 was also included. There were no considerable cell cycle dependent differences observed (Figure 4.4). After qPCR validation, the samples were sequenced. As a measure for sonication and sequencing biases, "input DNA" (cross-linked DNA fragmented under same conditions as the ChIPs) was used as appropriate control.

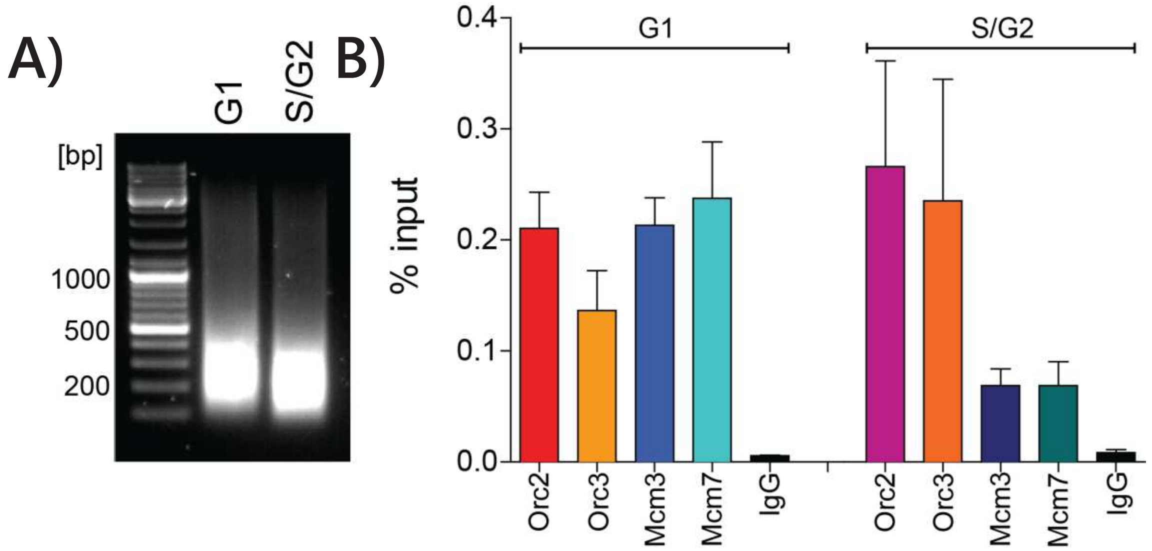


FIGURE 4.3: PRE-RC CHIP VALIDATION IN RAJI CELLS. A) REPRESENTATIVE EXAMPLE OF SONICATED CHROMATIN IN G1 AND S/G2. 1.5% agarose gel, stained with EtBr. B) PRE-RC CHIP VALIDATION BY QPCR AT THE DS LOCUS OF *ORIP.* Cell cycle stages as indicated. Rabbit IgG mock IP served as negative control. Mean ± SEM (n=3).

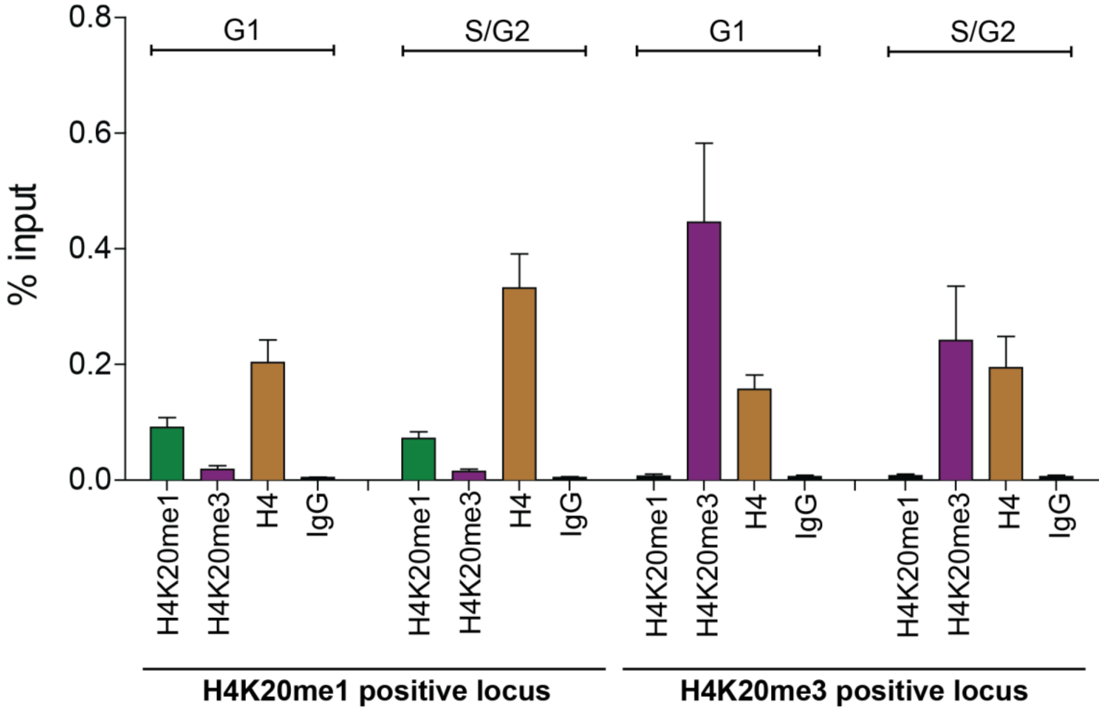


FIGURE 4.4: H4K20 METHYLATION CHIP VALIDATION BY QPCR AT GAPDH H4K20ME1 POSITIVE AND ZNF180 H4K20ME3 POSITIVE LOCI. Cell cycle stages as indicated. Rabbit IgG mock IP served as negative control. Mean ± SEM (n=3).

Treating sequencing data

The *Genomics unit of LAFUGA* (headed by Helmut Blum, LMU Munich) performed ChIP sample library preparations, as well as final 50bp single-end sequencing with the Illumina HiSEQ 1500 sequencer. Due to the expected variability of pre-RC ChIPs, I chose an elevated sequencing depth of around 70 million reads for pre-RC ChIPs and 35 million reads for histone modifications.

Quality of the ChIP-seq reads (or tags) was controlled using FastQC, and mapped against the human genome hg19 (GRCh37, version 2009), extended for the EBV genome (NC007605) using bowtie1. Choice of mapping parameters has an effect on sensitivity and specificity. Usage of only uniquely mapping reads excludes some true binding sites, simply because they locate in duplicated or repeated regions. However, allowing multiple read alignments leads to the detection of false positives. For the following analyses, false positives and possible amplification artefacts were excluded by blocking the repeated alignment of the identical sequence read.

An example of either input (input (G1) repl1) or Orc2-ChIP (Orc2 (G1) repl1) read sequences aligned on hg19 at the published origin Mcm4/PRKDC (Schaarschmidt *et al.* 2002) is shown in Appendix Figure 1 A. Visualization was performed with the Integrated Genome Browser (IGB). In principle, protein binding at a specific position leads to an accumulation of reads aligning at this locus. A direct comparison of sequencing alignments showing single reads and pileup profiles is depicted in Appendix Figure 1 B.

In general, raw ChIP-seq pileup profiles (without input normalization or any filters) turned out to be rather broad with much background, but I also observed regions with clear peaks (examples are visualized in Figure 4.14 (Mcm4/PRKDC locus, broad profile) and Figure 4.15 (chr6: 26516717-26543982, sharp peaks). Browser visualization is very useful to get an impression of the data, but is not quantitative for further analyses. For quantitative analysis, significant peaks need to be determined. Peak-calling programs compare read distribution of input to the specific ChIP and detect significant enrichments by applying dedicated algorithms.

My Laboratory previously determined positions of active replication and pre-RC on the EBV genome by SNS- or ChIP-on-chip experiments (Papior *et al.* 2012). Thereby, independent ChIP and SNS data sets displayed high concordance. To decide for the most appropriate peak-calling program, I applied different peak-calling algorithms on EBV ChIP-seq data and selected the program whose results matched best to the previously identified pre-RC and SNS positions on the EBV genome.

Validation of peak-calling on the EBV genome

Any peak-calling program's key task is to reproducibly identify correct protein binding positions while avoiding false positives. There are over fifty different peak calling programs available and I chose to compare three of them: i) the most popular one: MACS2 (Model-based analysis of ChIP-seq;

Zhang *et al.* 2008); ii) a similar stringent peak calling program with many downstream applications: HOMER (Hypergeometric optimization of motif enrichment; Heinz *et al.* 2010); and iii) a program that has been shown to detect an increased number of biological relevant peaks: T-PIC (Tree-shape peak identification for ChIP-seq; Hower, Evans, and Pachter 2011).

A ChIP-seq profile representation of single replicates and the respective peak calling result from EBV is depicted in Figure 4.5 for Orc2 and Orc3 and in Figure 4.6 for Mcm3 and Mcm7. While both MACS2 and HOMER peak calling programs did not detect any major peaks except for *oriP*, the T-PIC peak calling program also identified significant protein binding outside of *oriP* (Figure 4.5, Figure 4.6). This is consistent with published results, as Papior *et al.* found considerable pre-RC binding as well as active replication initiation events all over the EBV genome (dark brown (pre-RC) and light brown (SNS) bars in Figure 4.5, Figure 4.6; Papior *et al.* 2012).

These very divergent results of different peak-calling programs originate from the different underlying algorithms. Despite a lot of work invested, it remains difficult to accurately define a peak within the data. MACS2 is designated for transcription factor binding detection and histone modifications (Feng, Liu, and Zhang 2011). The principle of MACS2 peak-calling relies on a peak showing a bimodal enrichment pattern, as the sequencing reads represent each the end of a ChIP fragment (Figure 4.7). Based on the read distribution, MACS2 shifts the reads towards the 3'ends to optimize protein binding site location. Thereby, the shift size (d) is often unknown. MACS2 either relies on the sonication size (bandwidth parameter determined by the user) or calculates shift sizes based on high-quality peaks. Careful determination of the shift size is thus crucial for proper peak detection. MACS2 performs peak detection by using a dynamic parameter defined for each read enriched region (λ_{local}) and compares this enrichment to the peak region and neighboring regions (1 kb, 5 kb, 10 kb from the peak) in ChIP and control input sample. Thereby, variations in sequencing depth between ChIP and input samples are adjusted by diminishing the elevated depth. A p-value is calculated for each enriched region and those regions with a p-value smaller that the threshold (default $p=10^{-5}$) are considered significant (Zhang *et al.* 2008). MACS2 provides two general peak-calling functions: "standard" for punctuate enrichments as e.g. transcription factors and "broad" to detect enrichments from e.g. histone modification patterns that cover broader regions.

Also the HOMER "findPeaks" function allows choosing two different approaches. Either the setting "factor" for defined peaks, and "histone" for broader enriched regions can be chosen. To start with, HOMER requires to build a "tag directory" from the reads and performs a first quality control on data while estimating required parameters for downstream analysis (e.g. fragment length). Similar to MACS2, the program also shifts the enriched reads in the 3' direction to center the actual peak depending on the fragment size.

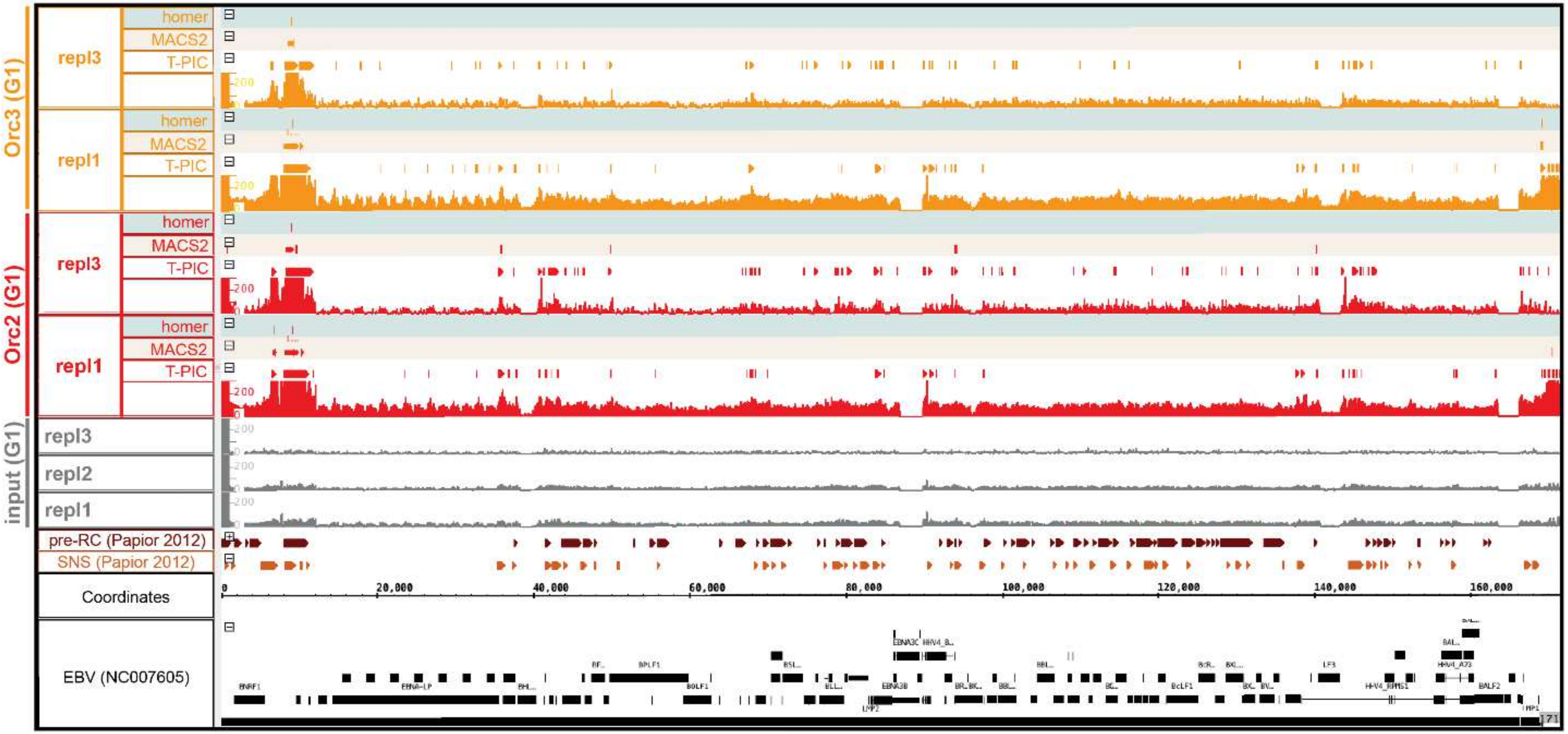


FIGURE 4.5: VALIDATION OF THE PEAK-CALLING ALGORITHM ON ORC2 AND ORC3 REPLICATES. Significant peaks represented as bars above each pileup profile, MACS2 peaks are colored in beige background, HOMER peaks in lightblue background. Input (grey), Orc2 (red), Orc3 (orange). Pre-RC (brown) and SNS (light brown) positions identified by Papior et al. 2012. EBV genes represented as black bars, black lines = introns. EBV NC007605 [chrEBV: 0-171823].

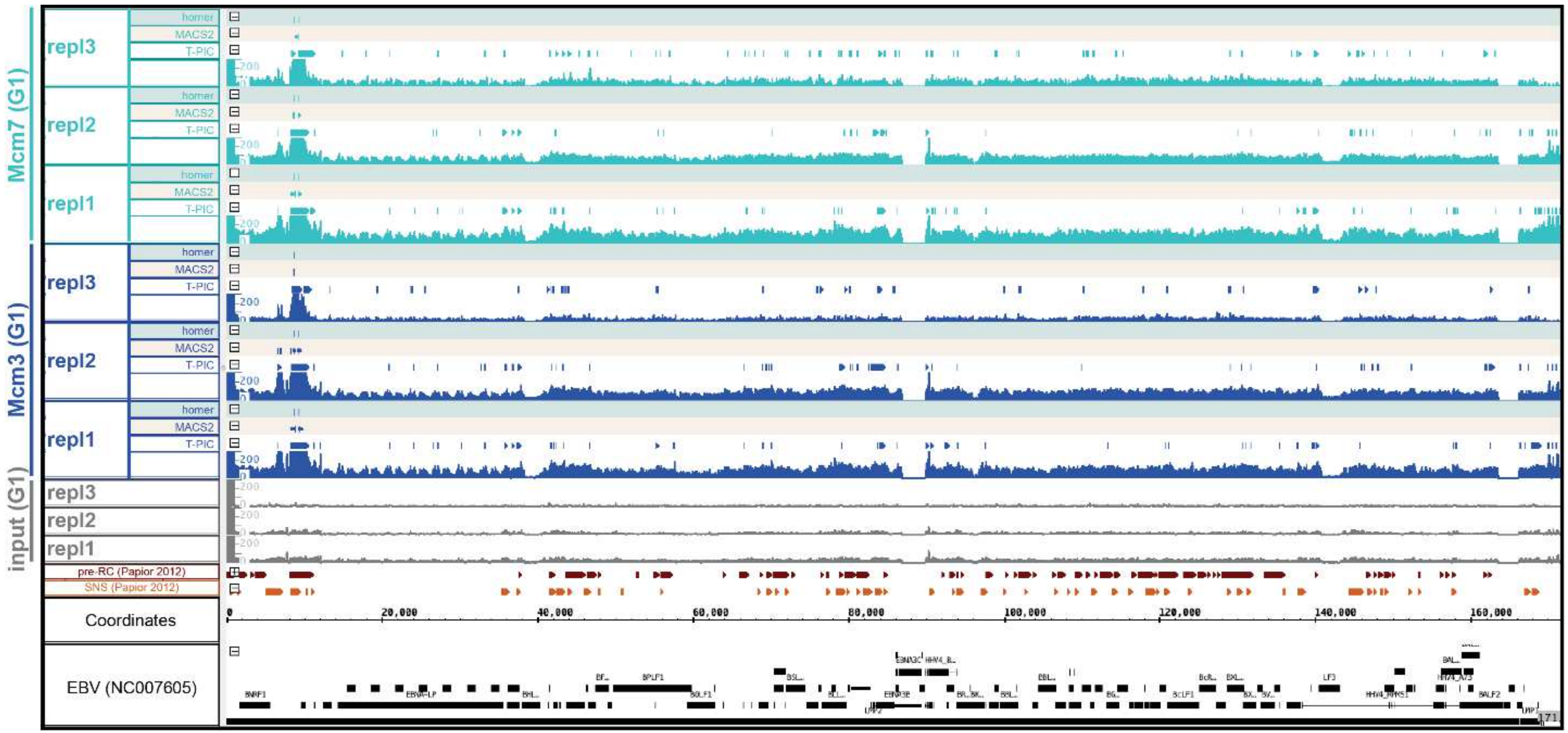


FIGURE 4.6: VALIDATION OF THE PEAK-CALLING ALGORITHM ON MCM3 AND MCM7 REPLICATES. Significant peaks represented as bars above each pileup profile, MACS2 peaks are colored in beige background, HOMER peaks in lightblue background. Input (grey), Mcm3 (blue), Mcm7 (turquois). Pre-RC (brown) and SNS (light brown) positions identified by Papior et al. 2012. EBV genes represented as black bars, black lines = introns. EBV NC007605 [chrEBV: 0-171823].

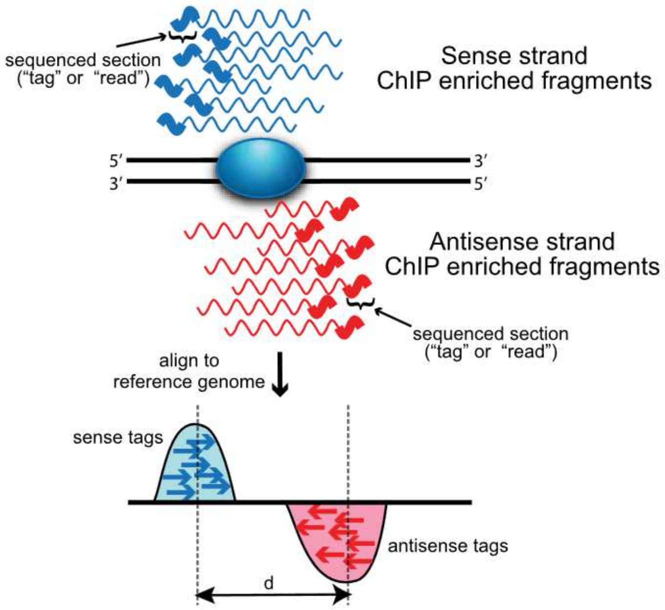


FIGURE 4.7: STRAND-DEPENDENT BIMODALITY IN SEQUENCING READS DISTRIBUTION. The protein of interest (blue oval) binds DNA. Wavy lines represent sense (blue) and antisense (red) DNA fragments from ChIP enrichments. Sequenced reads are indicated as thicker portion. Alignment to the reference genome produces bimodal distribution of a peak that is corrected by peak-calling programs. Distance between sense and antisense reads (d) represents optimal shift size. From Wilbanks and Facciotti 2010.

It scans the genome for clusters of high-density tags and excludes the regions directly adjacent to these clusters. Assuming that the local tag densities follow a Poisson distribution, the program estimates the expected peak numbers and simultaneously calculates the expected number of false positive peaks for each tag threshold. It then uses the threshold that meets the specified false discovery rate (default: fdr=0.001). Peaks are filtered against the input control experiment using the sequencing-depth independent fold-change parameter. This means that a potential peak in the target experiment needs 4-fold (default) more normalized tags than the control. Similar to MACS2, HOMER also filters peaks based on the local tag count. Peak tag densities have to be 4-fold above the surrounding 10 kb region (homer.salk.edu/homer/ngs/peaks.html).

In contrast to MACS2 or HOMER, T-PIC identifies significant peaks from read coverage and applies tree-based statistics on the data. The program starts by calculating the coverage from the reads of input and target sample extended by defined fragment length (set to 200 bp final). Already during this calculation, "anomalous" coverage is flagged for being subsequently analyzed as putative peaks. In order to detect statistically significant peaks, the program first calculates a "null hypothesis" to model regions without peak for a given coverage. The actual peak detection is then based on the shape of the peak (instead of it's simple height, also the information from the neighborhood is taken into account and used to differentiate from random fragment distributions). Potential peaks identified during coverage calculation are then reprocessed and the coverage change within the shape is represented as a "tree" (Figure 4.8). Sharp peaks consequently correspond to large trees without many branches, while broad peaks lead to low, rather bushy trees. A value is attributed to the tree which reflecting its shape. The "null hypothesis model" is equally scanned for peaks and corresponding trees are calculated to be able to correct for randomly occurring read accumulations. By comparing the detected peaks to the random peaks from the "null hypothesis model", significant peaks are identified and displayed (Hower, Evans, and Pachter 2011).

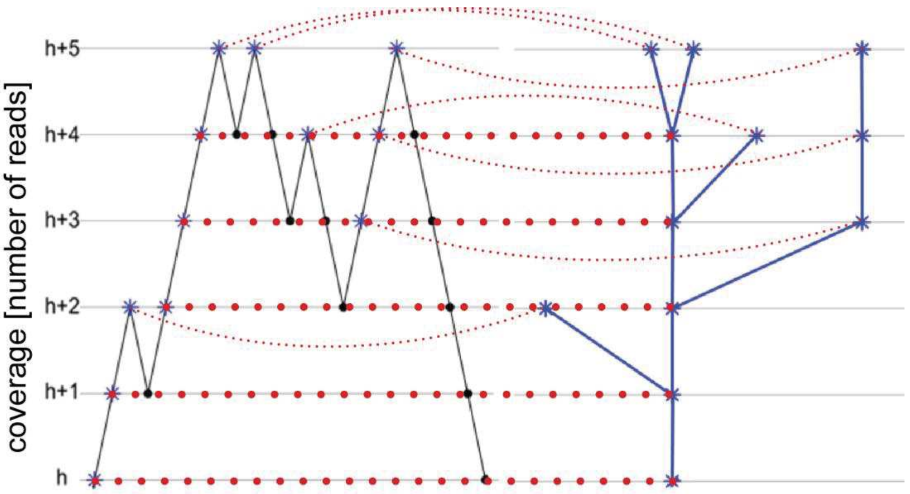


FIGURE 4.8: SCHEMATIC REPRESENTATION OF TRANSFORMATION OF A PEAK SHAPE INTO A TREE. The peak starts at a coverage height h and its end is considered at the same height. Every time the coverage changes by adding a read, the tree gains in vertices (blue stars, connected by red dotted line). The branches originate from "deviations" of the main shape (smaller dotted red lines). Adapted from Hower, Evans, and Pachter 2011.

This diverging mathematical models for peak-calling explain the differences of T-PIC to the other two peak-calling algorithms MACS2 and HOMER. Furthermore, the authors of the T-PIC algorithm validated their approach by re-analyzing already published data of factors with specific binding motifs. Indeed, T-PIC identified peaks that specifically contain the corresponding motifs, while other peak-calling programs (like MACS) missed some of these peaks (Hower, Evans, and Pachter 2011). In conclusion, T-PIC is a more sensitive program to detect peaks of biological relevance.

Based on the comparative analysis of the different peak-calling algorithms on the EBV pre-RC ChIP-seq results and the published EBV replication profile, T-PIC seems to be the most appropriate peak-calling program. I investigated the similarities of identified peaks between replicates and the different pre-RC proteins by calculating the Jaccard index. This index was calculated as explained in Figure 4.9 A. Jaccard index is a direct measure of replicate similarities with 1 being exactly redundant and 0 being completely different. Jaccard indices can be represented by a heatmap, which allows direct visual estimation of ChIP reproducibility. As observed in Figure 4.9 B, ChIP peaks clustered principally by experiment. Replicate 1, 2, and 3 cluster among themselves, but there was also a clustering visible depending on the ORC or Mcm2-7 complex. ChIPs from the third experiment (repl3) were least similar.

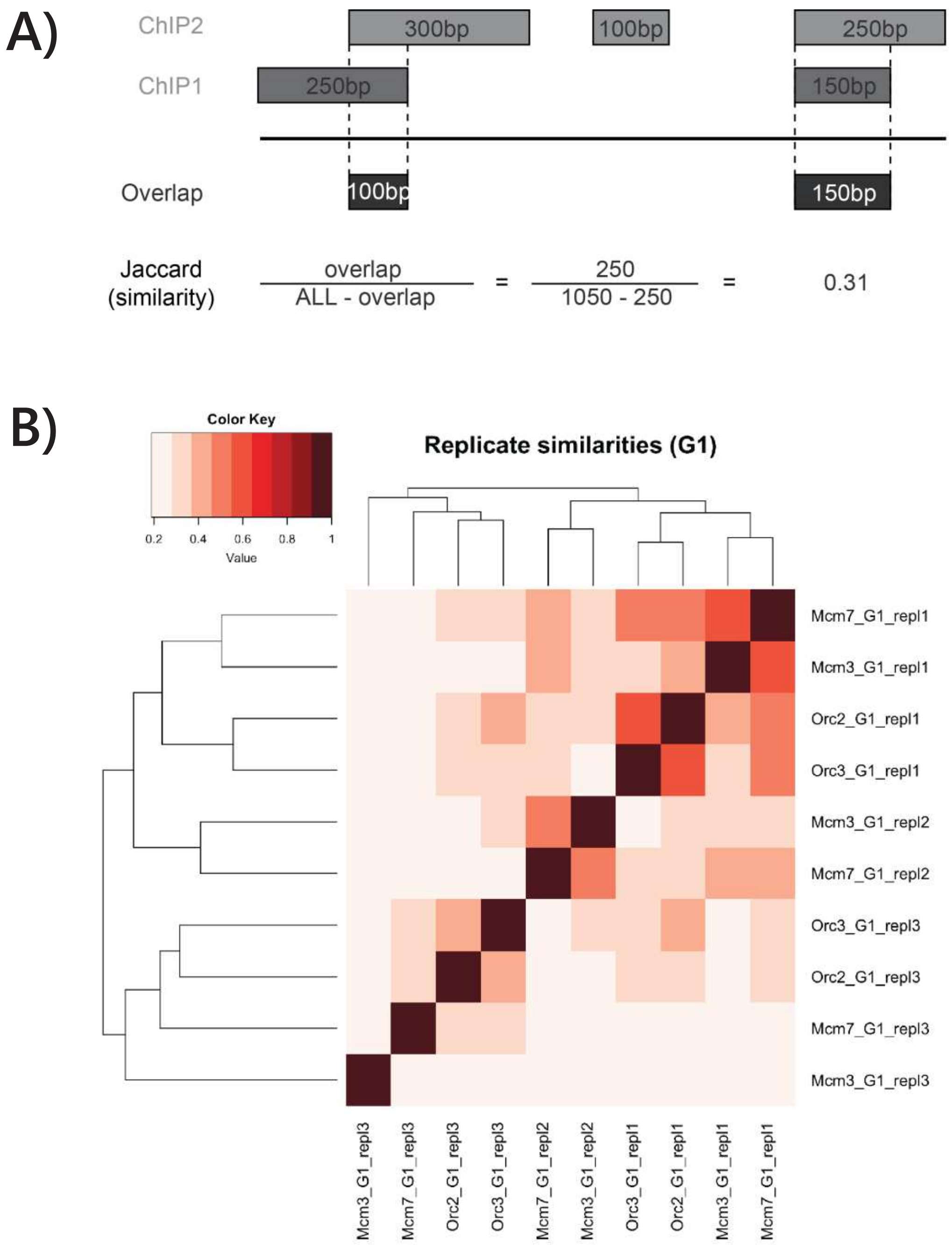


FIGURE 4.9: CHIP SIMILARITIES REPRESENTED BY JACCARD INDICES. A) SCHEMATIC EXPLANATION OF CALCULATING THE JACCARD INDEX. The peak properties of two ChIPs are directly compared by calculating the overlap of both peaks and dividing by the sum of the totality and the overlap. The resulting index is a direct measure of similarity and ranges between 0 (no similarity) and 1 (identical). B) HEATMAP REPRESENTATION OF JACCARD INDICES OF THE INDIVIDUAL CHIP T-PIC PEAKS ON THE EBV GENOME. Dark red represents great similarities while lighter red indicates less redundancies.

These observed differences resulted principally from the biological and experimental difference of the samples. Working with three independent elutriation fractions, that were cross-linked at different time-points might account for experimental differences. Also, Mcm2-7 complexes are not necessarily positioned at defined sites but slide away from their ORC-dependent loading sites (discussed in more detail in chapter 4.3). This might explain the increased variance in Mcm2-7 protein ChIPs.

DEFINING THE PRE-REPLICATION COMPLEX

The high experimental and biological variances strengthen even more the necessity for stringent pre-RC definition. The availability of several replicates (two for each Orc-ChIP and three for each Mcm2-7 ChIP) and two target proteins within each pre-RC sub-complex, makes robust complex definitions possible. Because ORC can also exist without Mcm2-7 helicases and the Mcm2-7 helicases might define initiation sites on their own, I defined and examined three different complexes: ORC, Mcm2-7, and pre-RC. A complex is per definition composed of at least 50% overlapping peaks within respective target ChIPs and replicates (Figure 4.10 A). Consequently, ORC consisted of at least two peaks overlapping from Orc2_repl1, Orc2_repl3, Orc3_repl1, Orc3_repl3, Mcm2-7 demanded at least three overlapping peaks in Mcm3_repl1, Mcm3_repl2, Mcm3_repl3, Mcm7_repl1, Mcm7_repl2, and Mcm7_repl3. Pre-RC was built of at least five overlaps of all ChIPs.

A visualization of the determined T-PIC-defined complex distributions together with pre-RC and SNS determined by Papior *et al.* 2012 is shown in Figure 4.10 C. W-repeats and terminal-repeats were excluded from the analysis. When calculating overlaps of T-PIC-defined pre-RCs with pre-RC determined by Papior *et al.*, 66% of all T-PIC-pre-RCs overlapped. The same was true regarding the overlap of T-PIC-pre-RCs with SNS (Figure 4.10 B). Interestingly, one particular region from 102000 to 137600 was heavily enriched for pre-RCs detected by Papior *et al.*, while T-PIC did not detect any enrichments of ORC, Mcm2-7 or pre-RC. Either this region is constantly covered by pre-RC components, which impairs T-PIC from detecting significant enrichments, or this observation results from differences in technique and bioinformatics analysis.

Jaccard indices of all determined complexes were also calculated and are represented in Appendix Figure 2. T-PIC-defined complexes clustered together with high similarities and were slightly more similar to SNS positions determined by Papior *et al.* (Jaccard index = 0.32) than to their pre-RC positions (Jaccard index = 0.18).

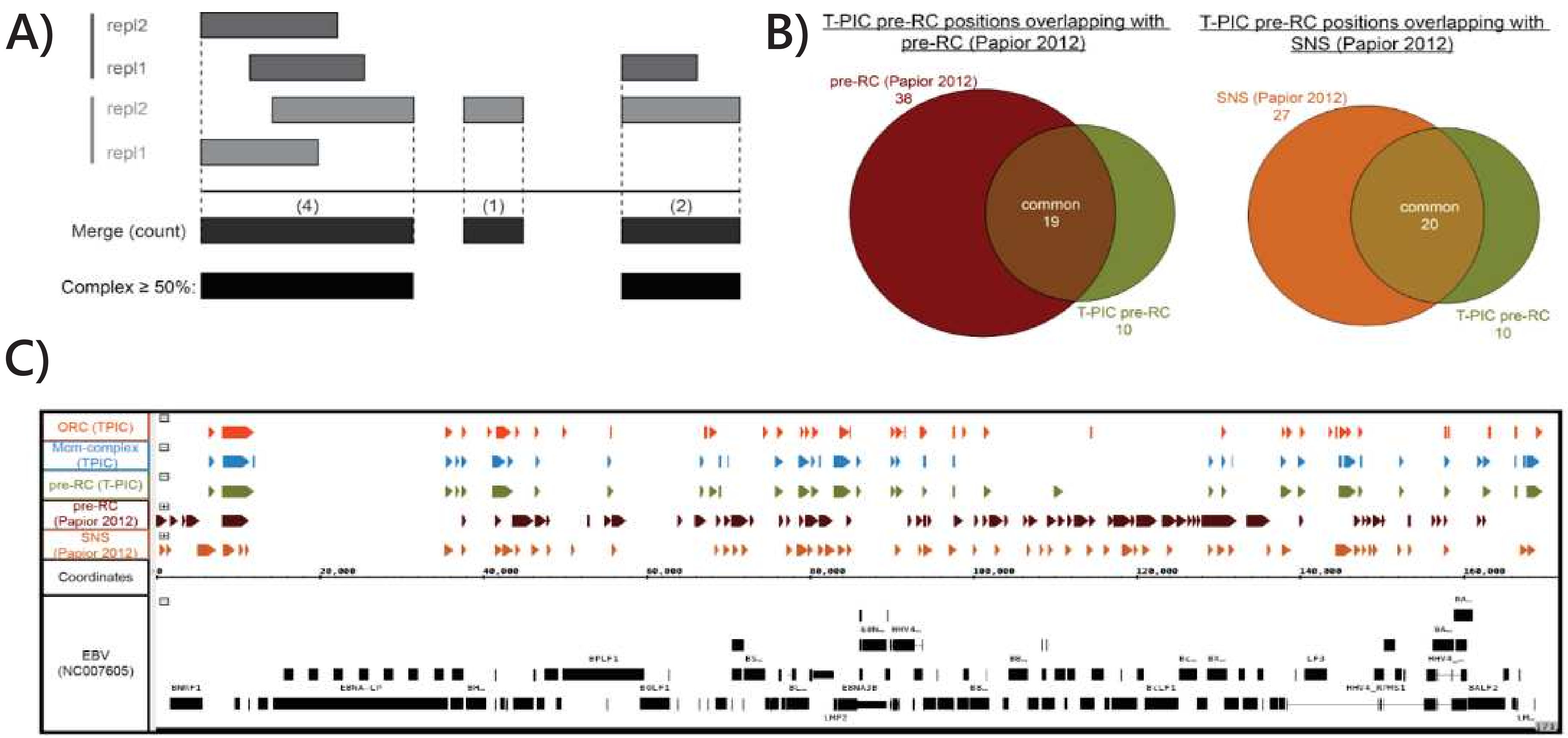


FIGURE 4.10: ORC, MCM2-7, AND PRE-RC DIFINITION ON EBV GENOME. A) PRINCIPLE OF COMPLEX DEFINITION. At least 50% overlapping peaks are needed to determine a complex. During merging, leftmost start position and rightmost end position are conserved. B) DIRECT COMPARISON OF T-PIC DEFINED PRE-RC POSITIONS WITH PRE-RC AND SNS POSITIONS FROM PAPIOR *ET AL.* Pre-RC (Papior et al. 2012) are represented in brown, SNS (Papior et al. 2012) in light brown and T-PIC-defined pre-RCs in green. C) IGB VISUALIZATION OF T-PIC-DEFINED ORC, MCM2-7 AND PRE-RC, TOGETHER WITH PRE-RC AND SNS DETERMINED BY PAPIOR *ET AL.* 2012 ON THE EBV GENOME. T-PIC ORC: orange, T-PIC Mcm2-7: blue, T-PIC pre-RC: green, pre-RC (Papior *et al.* 2012): brown, SNS (Papior *et al.* 2012): lightbrown; [chrEBV: 0-171823].

In conclusion, the T-PIC peak-calling algorithm is most suitable to detect relevant peaks from pre-RC ChIPs. The observed differences of T-PIC-defined peaks to pre-RC and SNS positions defined by Papior *et al.* might account for completely different technical and bioinformatical approaches. Even more, the fact that I also observed considerable similarities (66% overlaps) argues for my chosen analysis procedure.

4.1.3 PERFORMING CHIP-SEQ EXPERIMENTS IN HUMAN EMBRYONIC STEM (HES) CELLS

Setting up pre-RC ChIP-seq experiments in Raji cells had the advantage that an internal ChIP quality control was available with *oriP* on the EBV genome. The resulting sequencing data now allowed determination of genomic pre-RC positive loci to enable qPCR validation of pre-RC ChIPs in other human cells. Together with my partner laboratory of Dr. Jean-Marc Lemaitre at the IRMB in Montpellier, we decided to apply pre-RC ChIP-seq on the human embryonic stem cell line H9 (hES cells). In these cells, SNS-sequencing data is also available, which allows direct comparison of pre-RC positions with specific sites of active replication (Besnard *et al.* 2012). However, easy G1 synchronization of hES cells is impossible and I decided against chemical synchronization to avoid treatment biases, performing the ChIPs on asynchronous cell populations.

EXPERIMENTAL SETUP: CHIP VALIDATION

ES cells were cultivated to high quantities (up to $2x10^8$ cells). This made careful pluripotency controls necessary, as hES cells easily differentiate. Besides morphological criteria (Appendix Figure 3 A), I also performed FACS staining of known pluripotency markers SSEA4 and Oct4 (Appendix Figure 3 B). FACS stain revealed a pluripotent population of about 98% of all cells. During cross-link using my established protocol, I omitted the second washing step prior to cross-linking because cells were more fragile and lysed more easily. Sonication time was also reduced and ChIPs were performed as described in chapter 3.2.7, p. 35.

From Raji sequencing data, I chose two positive loci for ChIP validation in hES cells by qPCR. The Mcm4/PRKDC locus represents a known origin of replication (Schaarschmidt *et al.* 2002) and primer positions together with example ChIP-seq profiles in Raji cells are shown in Appendix Figure 4. When screening Raji ChIP-seq data for strong enrichment peaks, I defined the BANF1 locus as additional positive control (sequencing profile and primer positions are shown in Appendix Figure 5). After validation of positive and negative control primers in Raji cells (not shown), I performed pre-RC ChIPs according to the established protocol and validated ChIP quality by qPCR (Figure 4.11).

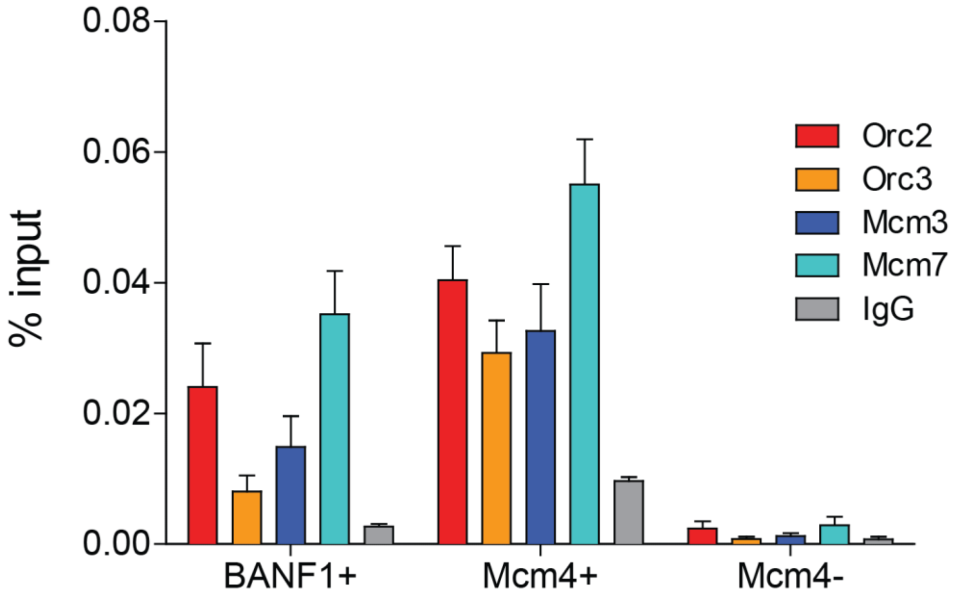


FIGURE 4.11: PRE-RC CHIP QPCR VALIDATION AT SELECTED BANF1 (BANF1+) AND MCM4/PRKDC POSITIVE (MCM4+) AND MCM4/PRKDC NEGATIVE (MCM4-) LOCI. IgG was used as negative control. Mean ± SEM (n=3).

Pre-RC ChIP validation on the human genome resulted in low % input values, which possibly emerge from the variable binding of pre-RC proteins. However, ChIPs remained enriched at positive loci in comparison to the Mcm4/PRKDC negative locus and were sequenced at the platform MGX-Montpellier GenomiX.

Two complete replicates of Orc2, Orc3, Mcm3, and Mcm7 were sequenced by paired-end sequencing of 100 nucleotides from both ends of the ChIP fragments. This method improves data set quality because alignment to the reference genome is facilitated and also repetitive sequence elements can be covered.

TREATING SEQUENCING DATA

Anissa Zouaoui, bioinformatician in the laboratory of Dr. Jean-Marc Lemaitre, mapped the sequencing data against the human genome hg19 (GRCh37, version 2009) and performed standard MACS2 broad peak calling, as pre-RC binding turned out to be more spread over the genome, rather than representing distinct peaks. Currently, we are also adapting the T-PIC peak calling algorithm for paired-end sequencing data, as T-PIC is originally designed for single-end sequencing results. Nevertheless, to get a first impression of the data, I defined ORC, Mcm2-7 and pre-RC as previously described and compared their positions to sites of active replication (SNS, Besnard *et al.* 2012).

First of all, I looked at the total number of defined complexes from replicates (Figure 4.12 A). It is evident that many more Mcm2-7 complexes (43846) were detected than ORC (12848) or pre-RCs (8651). By extensive deep sequencing, Besnard *et al.* 2012 detected more than 200000 sites of active replication. Although this elevated number of origins has been recently questioned by Cayrou *et al.* 2015 and highly depends on definition and detection algorithm, it remained very robust when changing algorithm parameters (data not shown). Evidently, the number of defined ORC, Mcm2-7 and pre-RCs was smaller and cannot be directly compared to the number of active origins. Nevertheless, I analyzed the overlaps of the defined complexes with SNS (Figure 4.12 B). While ORC and pre-RC overlapped to nearly 80% with SNS sites, only 30% of all Mcm2-7 complexes were found at SNS sites.

Given that mainly ORC and pre-RC complexes overlapped with SNS, I wondered for the distance of all the complexes towards SNS center. I calculated and plotted the distance of the defined complexes towards the next SNS center in log10 scale in a bar chart (Figure 4.12 C). Distribution of close and more farther located complex populations gives visual impression of the behavior towards SNS.

It becomes evident that majority of ORC (80.5%) and pre-RC (80.3%) located within ~1600 bp (log10 = 3.2, marked by red dashed line in Figure 4.12 C) of SNS center, while only 40.4% of Mcm2-7 located that close. This also corresponded to the directly calculated overlaps and shows again, that while ORC was located in closer proximity to SNS (mean=4561 bp), Mcm2-7 complexes were detected further away (22052 bp, $p < 2.2x10^{-16}$).

The observed 3.4-fold excess of the total Mcm2-7 numbers of compared to ORC is consistent with already reported Mcm2-7 helicase surplus in *Drosophila*, although not to the same extend (Powell *et al.* 2015 reported 10 to 50-fold Mcm2-7 helicase excess). The number of detected ORC (12848) corresponds to the number of Orc1 sites detected by Dellino *et al.* 2013 (~13000) in HeLa cells, but not to ~52000 Orc2 peaks in K562 cells (Miotto, Ji, and Struhl 2016) neither to the expected number of total origins (30000-50000). Although the numbers and positions of ORC and Mcm2-7 differ considerably, 67% of all defined ORC coincide with Mcm2-7, which resulted in 8651 detected pre-RC positions. Apart from apparent technical fluctuations in antibody specificities, this might reflect ORC-dependent Mcm2-7 loading (overlapping ORC/Mcm2-7) and subsequent Mcm2-7 sliding from the loading site (Mcm2-7 without corresponding ORC), as suggested previously (Das and Rhind 2016; Hyrien 2016).

The strong overlap of ORC, but not Mcm2-7 with SNS contradicts the idea that Mcm2-7 complexes are solely responsible for replication initiation. In that case, mostly Mcm2-7 should overlap with replication initiation sites. These results rather indicate, that ORC is indeed important defining active origins, and it would be interesting to test, which cohort of origins overlaps with ORC (in terms of origin efficiency, links to transcription or histone modifications, etc.). However, it would also be interesting to further define the Mcm2-7 population not correlating with SNS, to find possible explanations.

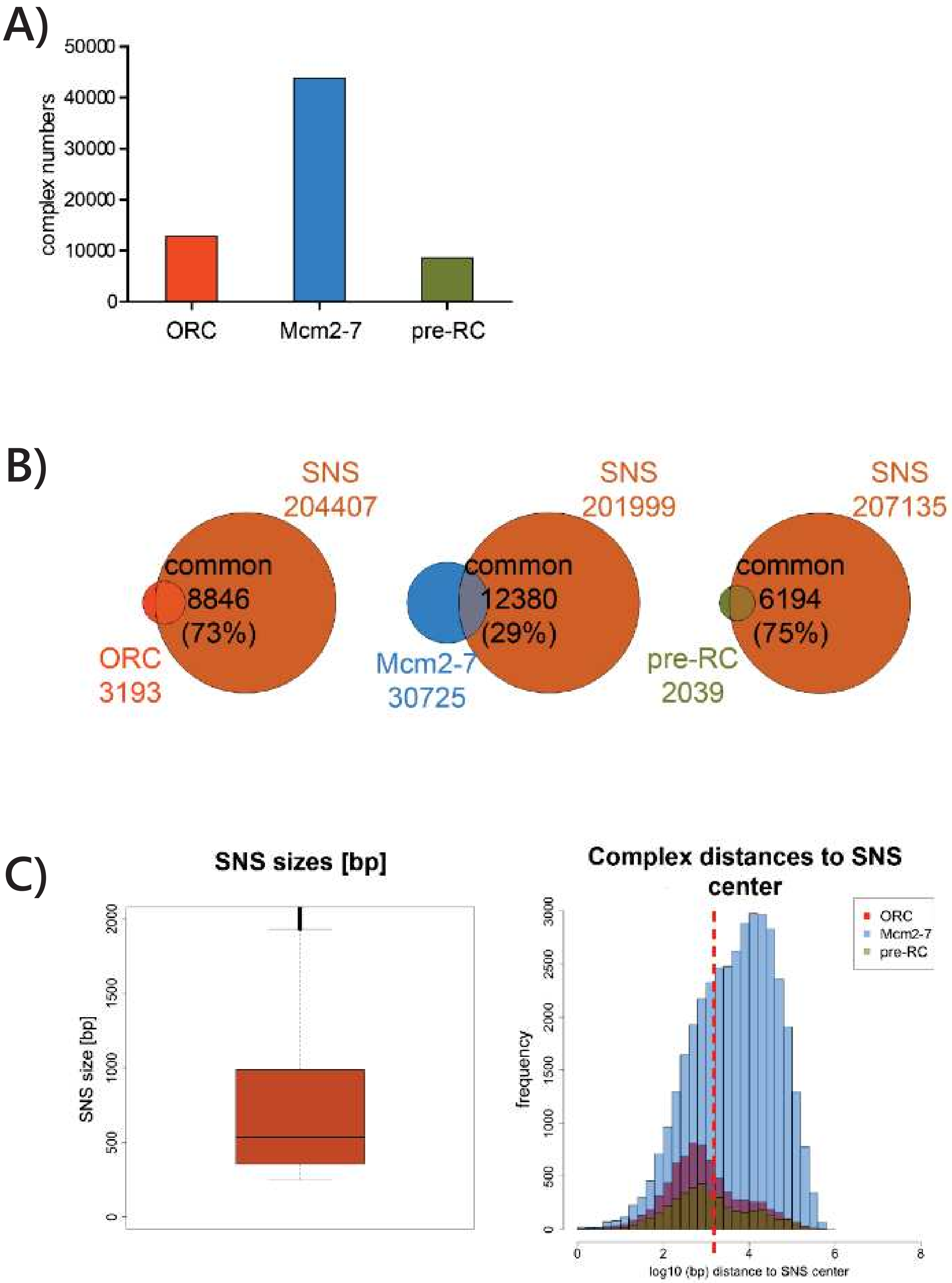


FIGURE 4.12: SNS CORRELATE MORE WITH ORC THAN WITH MCM2-7. A) TOTAL NUMBERS OF DEFINED COMPLEXES. B) OVERLAP OF DEFINED COMPLEXES WITH SNS. Percentages in brackets represent the % overlap of the complex with SNS. C) SNS SIZE DISTRIBUTION AND DISTANCE TO SNS CENTER ANALYSIS OF EACH DEFINED COMPLEX. The frequency of the distance of each complex was calculated in respect to the center of the next SNS position and plotted as log10. The red dashed line marks log10(3.2) = 1585bp from center. ORC: orange, Mcm2-7: blue, pre-RC: green, SNS: lightbrown.

In conclusion, although the data seems promising, it still needs extensive further analysis in collaboration with Anissa Zouaoui. First effort will be to apply the earlier validated T-PIC peak calling algorithm on the data to gain confidence in the peak number and positions detected. This will also improve the comparative analysis with SNS positions. Further steps will include the correlation of SNS and pre-RC positions with other chromatin features, like histone modifications, MNase sensitivity, CG-content, etc. This will allow to define features predicting pre-RC correlation with SNS.

4.2 PRE-RC COMPONENTS ARE ENRICHED WITHIN REPLICATION UNITS

Pre-RC ChIP-seq analysis in hES cells still needs further effort concerning data treatment and evaluation. Besides comparison to active replication, there is a lot of histone modification ChIP-seq data available in hES cells and comparative correlations promise insights into regulation of origin licensing and activation.

For Raji pre-RC ChIP-seq comparisons however, there is less data published, and I initiated collaborations with Dr. Olivier Hyrien (IBENS, Paris) for active replication data, Prof. Dr. Wolfgang Hammerschmidt (HelmholtzZentrum München) for RNA-seq data, and Dr. Jean-Christophe Andrau (IGMM, Montpellier) for histone modification data.

To start data analysis on the human genome, I performed T-PIC and HOMER peak-calling on pre-RC ChIP-seq data sets. T-PIC is the peak calling algorithm of choice, as validated on the EBV genome. However, I decided to also separately focus exclusively on the strongest pre-RC peaks and to in parallel use the HOMER algorithm.

4.2.1 COMPARING T-PIC VS. HOMER PEAK-CALLING ALGORITHMS AND DEFINING ORC, MCM2-7, AND PRE-RC ON THE HUMAN GENOME

As already observed on the EBV genome, T-PIC peak-calling is much more sensitive than the HOMER algorithm. This was also true for peak-calling on the human genome. As a consequence, peak-calling with T-PIC detected around 20-times more peaks than HOMER (Table 4.1). Most HOMER peaks resided within the top 10% of T-PIC peaks (representative example for one replicate of Orc2 and Mcm3 depicted in Figure 4.13, see Appendix Figure 6 for Orc3 and Mcm7), confirming that HOMER mostly calls the strongest peaks.

TABLE 4.1: COMPARISON OF ABSOLUTE NUMBER OF PEAKS WITHIN REPLICATES AND RESULTING COMPLEXES. n.a = not assessed

		PEAK NUMBERS				COMPLEX NUMBERS		
		Orc2	Orc3	Mcm3	Mcm7	ORC	Mcm2-7	Pre-RC
HOMER	repl1	2309	4503	2203	2584	1936	322	329
	repl2	n.a.	n.a.	658	767			
	repl3	1391	1997	424	405			
T-PIC	repl1	52366	63976	76032	50795	57597	25529	23896
	repl2	n.a.	n.a.	44400	25060			
	repl3	84343	72717	53792	57388			

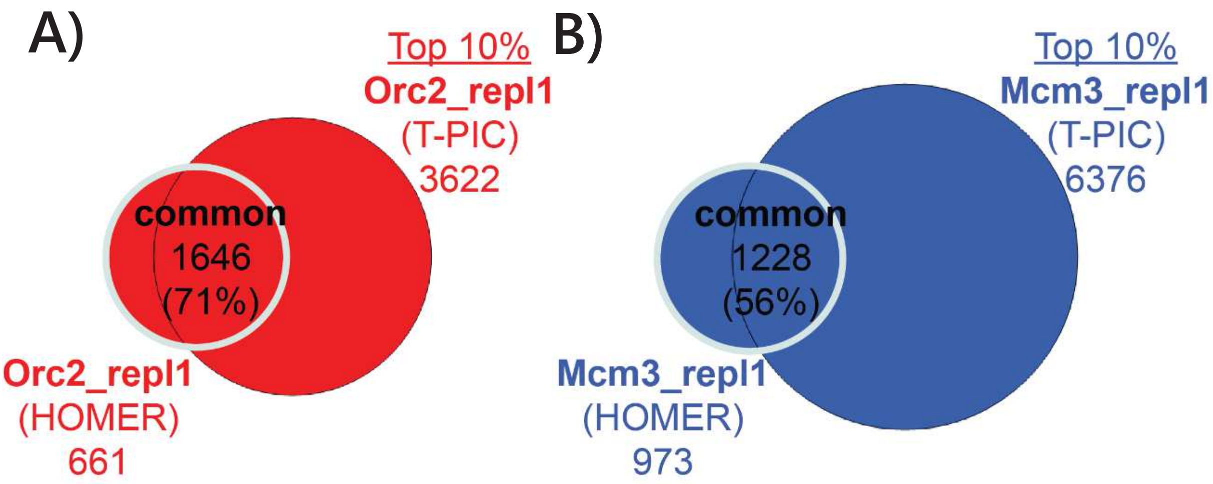


FIGURE 4.13: MOST HOMER PEAKS LOCATE WITHIN THE TOP 10% OF T-PIC PEAKS. REPRESENTATIVE EXAMPLE ANALYSIS FOR A) ORC2 AND B) MCM3 IN ONE REPLICATE. Venn diagram of overlap between all HOMER-detected peaks with the top 10% of T-PIC-detected peaks. Overall counts are indicated. The percentage of overlapping HOMER-detected peaks are specified in brackets.

Due to its high sensitivity, T-PIC detected peaks at many known origins of replication while HOMER did not (for Mcm4/PRKDC locus, see Figure 4.14, for LaminB2 (Abdurashidova *et al.* 2000) and JunB (Fu *et al.* 2014) origins, Appendix Figure 7 and Appendix Figure 8). These observations again confirm the T-PIC program detecting peaks of biological relevance. By contrast, HOMER mostly detected strong, highly enriched binding sites (example in Figure 4.15).

The number of detected peaks varied considerably between the different replicates (Table 4.1), emphasizing even more the necessity for stringent complex definition. On the basis of T-PIC and HOMER peaks, I computed ORC, Mcm2-7 and pre-RC according to the settings previously defined on EBV. The final complex numbers are listed in Table 4.1. For both approaches, the total number of ORC was 2 to 6-fold higher than both Mcm2-7 and pre-RC. This stands in clear contrast to the 3.4-fold excess of Mcm2-7 detected in hES cells and likely depends on the “broad” peak-calling performed with MACS2 in hES cells. HOMER-defined complex numbers were generally 30 to 80-times lower than T-PIC defined complexes.

Comparison of complex sizes revealed that HOMER complexes were considerably narrower than T-PIC ones (for pre-RC, see Figure 4.16 A, ORC an Mcm2-7 in Appendix Figure 9). This is mostly due to the HOMER peak-calling algorithm settings that produced peaks of around 200 bp. T-PIC peak sizes were more variable, also due to the merging process during complex computation.

When looking at the distribution of ORC, Mcm2-7 and pre-RC complexes, it became evident that HOMER-defined complexes represent a subpopulation of T-PIC-defined complexes (Figure 4.16 B (pre-RC), Appendix Figure 10 (ORC and Mcm2-7)). This is a logic consequence of HOMER peaks being part of the strongest T-PIC peak population but also reflects the fact that HOMER-defined complexes (and consequently the ones with very strong binding) only built a small sup-population within T-PIC-defined complexes.

After defining ORC, Mcm2-7 complex and pre-RC, the complexes can now be correlated with active replication, transcription and histone modifications.

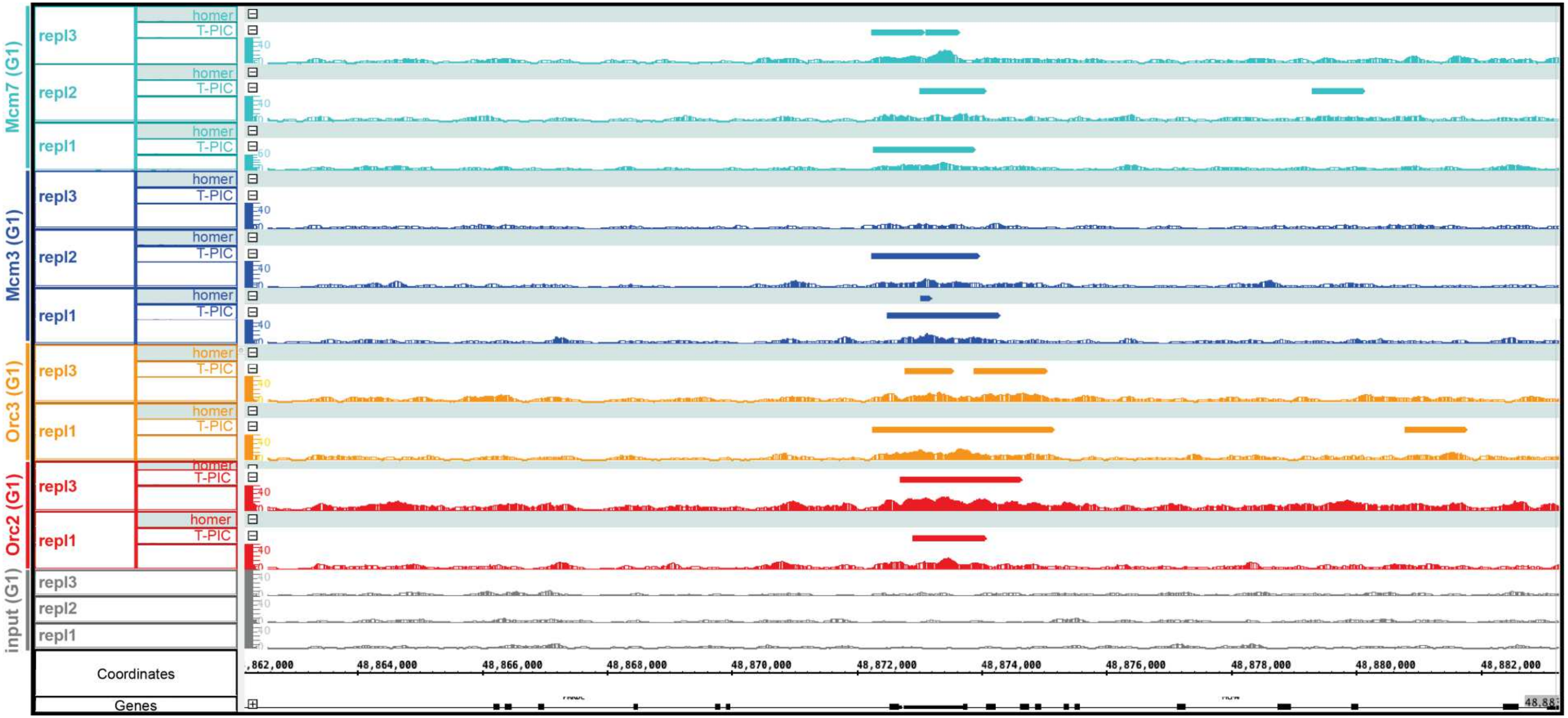


FIGURE 4.14: MCM4/PRKDC ORIGIN IS IDENTIFIED BY T-PIC BUT NOT BY HOMER PEAK-CALLING. Input: grey, Orc2: red, Orc3: orange, Mcm3: blue, Mcm7: turquois. T-PIC-detected peaks are represented as bars above each pileup profile; HOMER-detected peaks are additionally highlighted by lightblue background. Visualization in IGB. [chr8: 48862188-48883250].

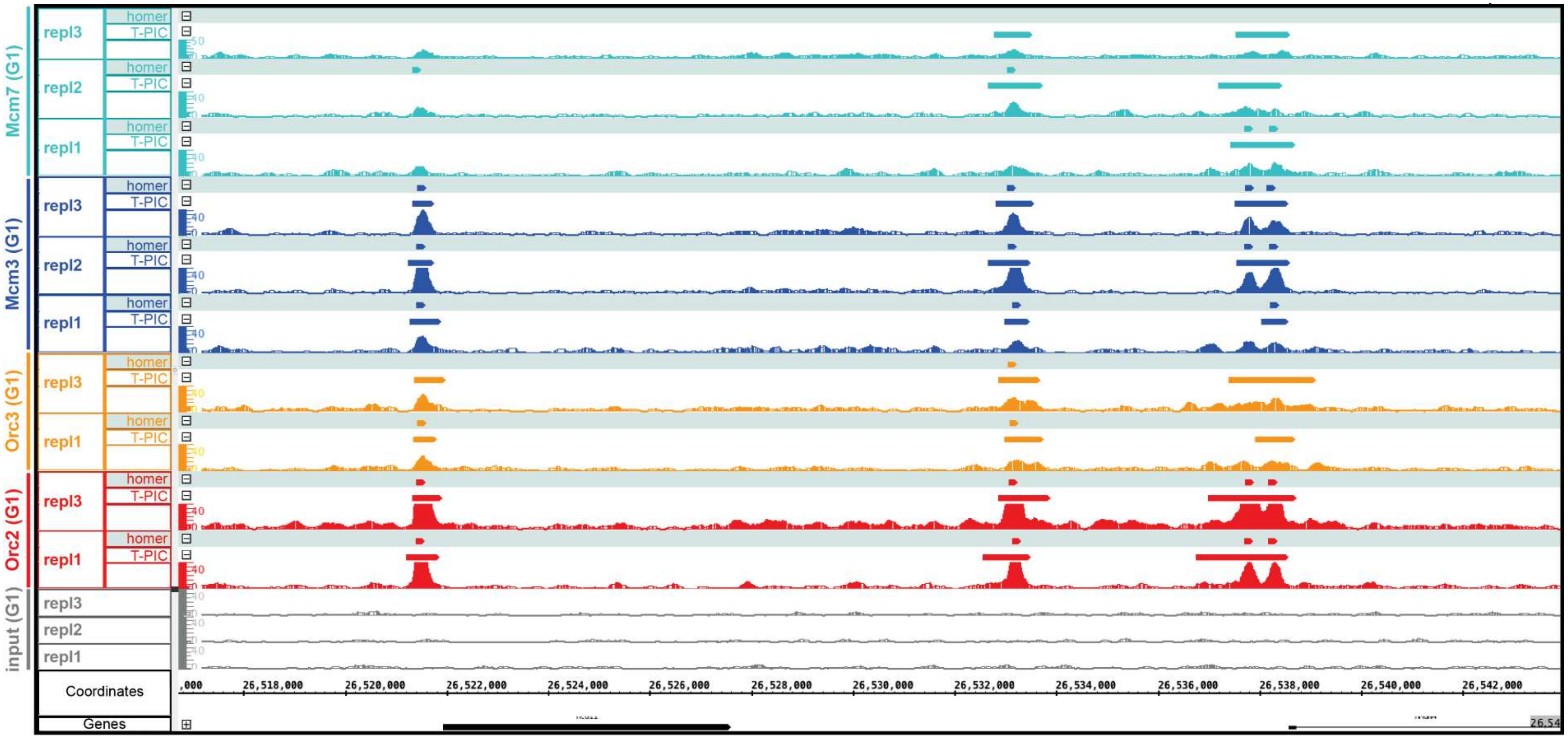


FIGURE 4.15: HOMER PEAK-CALLING IDENTIFIES STRONG PEAKS. Example peak-calling at chr6: 26516717-26543982. Input: grey, Orc2: red, Orc3: orange, Mcm3: blue, Mcm7: turquois. T-PIC-detected peaks are represented as bars above each pileup profile; HOMER-detected peaks are additionally highlighted by lightblue background. Visualized in IGB.

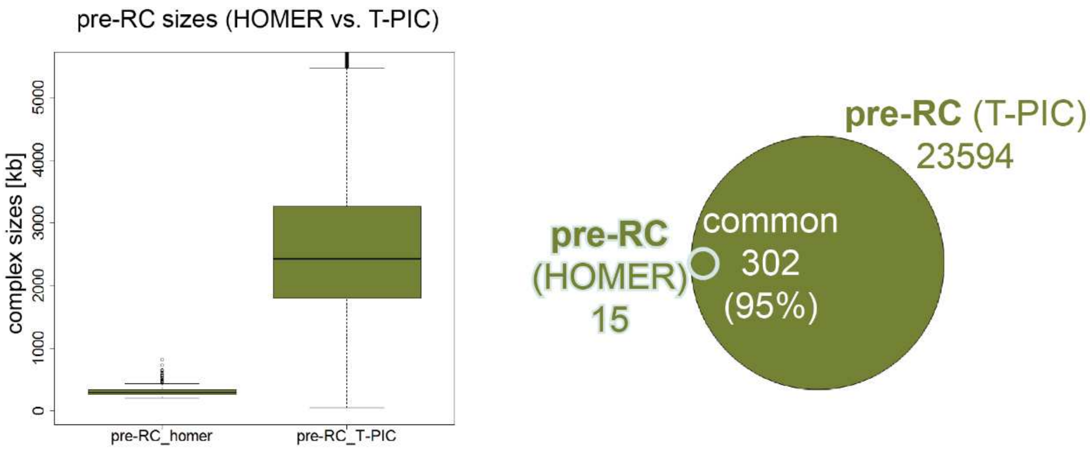


FIGURE 4.16: DIRECT COMPARISON OF HOMER- AND T-PIC-DEFINED PRE-RC. A) PRE-RC SIZES IN [KB]. Represented in boxplot: thick line shows the median, the box is the distribution from the first to the third quartile, the whiskers indicate the smallest and largest value without being an outlier. B) OVERLAPPING PRE-RCS. Venn diagram of overlap between HOMER- and T-PIC-defined pre-RCs. Overall counts are indicated. The percentage of HOMER-defined pre-RCs common with T-PIC-defined pre-RCs is specified in brackets.

4.2.2 PRE-RCS CORRELATE WITH ACTIVE REPLICATION UNITS

As pre-RCs represent sites of licensed origins of replication, I first assessed the association of origin licensing and active replication. The group of Dr. Olivier Hyrien adopted very recently a method for Okazaki-fragment sequencing (OK-seq) on human cells to detect replication initiation (Petryk *et al.* 2016). They also applied this method on Raji cells and kindly provided preliminary data on replication initiation and replication termination zones.

Principal of OK-seq is pulse-labeling of Okazaki-fragments with the Thymidine analogue EdU and subsequent purification of these fragments (< 200 bp). Fragment sequencing allows to distinguish between Watson (leftward moving fork) and Crick (rightward moving fork) Okazaki fragments. From this differentiation, replication fork directionality (RFD = $\frac{(Crick-Watson)}{(Crick+Watson)}$) is calculated resulting in a profile of series of ascending (AS), descending (DS), and flat segments of different sizes and slopes (Figure 4.17 A). AS represent zones of preferential replication initiation, while DS are zones of preferential replication termination. The amplitude thereby reflects initiation efficiency. A broad initiation zone (AS) represents a zone of preferential replication initiation with multiple inefficient initiation sites but only one single origin firing per cell, presumably corresponding to replication units described in the introduction (chapter 1.1.5, p. 15f). Computational wavelet detection of AS and DS in Raji cells (performed by Benjamin Audit, ENS Lyon) revealed 4639 AS and 2207 DS, covering 6.6% and 8.6% of the genome, respectively. Thereby, AS were smaller (mean = 37.8 kb, range from 3.9 to 198.0 kb) than DS (mean = 122,2 kb, range from 4.7 kb to 339.8 kb) (Figure 4.17 C).

Visual examination of AS, DS and T-PIC-defined pre-RC positions reveals a pre-RC enrichment in AS

First visual examination of the co-occurrence of AS, DS and T-PIC-defined pre-RC positions revealed a preferential localization of T-PIC pre-RCs within AS, while there were less pre-RCs found in DS (IGB visualized example in Figure 4.17 B). To quantify this visual observation, I calculated the density of T-PIC pre-RC positions per Mb (Figure 4.17 D). While 8.5 events/Mb were observed in AS, only 3.4 events/Mb were detected in DS. Additionally, I also considered all genomic regions being neither AS nor DS, termed "remaining genome". With 8 events/Mb, pre-RC density was similar to AS. Interestingly, these results do not indicate preferential pre-RC positioning within AS zones of active replication initiation, but rather argue for a specific depletion of pre-RCs in DS zones of replication termination.

On average, 8.5 pre-RCs per Mb were detected within an AS, which results in one pre-RC every 100 kb. Concordantly, on average, one origin is activated every 100 kb (mean size of a replication unit, Fragkos *et al.* 2015). However, one would expect to also detect the excess of licensed origins, thus anticipating a higher pre-RC density. This might possibly be a problem of peak detection, as every peak-calling algorithm expects protein binding at a defined region. This prerequisite is given for ORC but for Mcm2-7 complexes site specific binding is controversially discussed. Assuming Mcm2-7 spreading, Mcm2-7 binding sites might remain undetected. Consequently, I chose another approach to unbiasedly detect pre-RC omponent binding by calculating the mean read coverage of every pre-RC component at AS or DS.

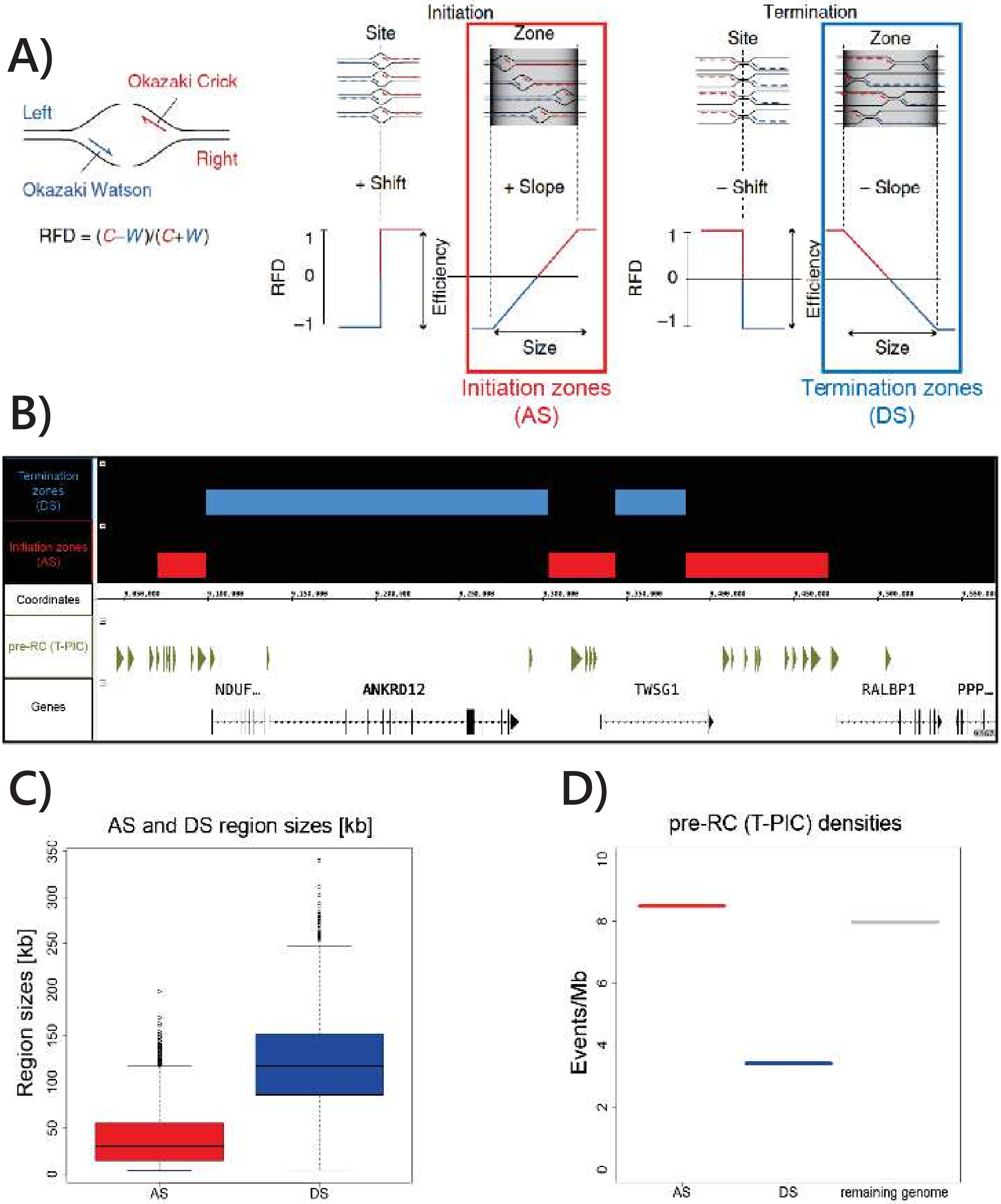


FIGURE 4.17: CORRELATION OF PRE-RC PEAKS WITH ZONES OF REPLICATION INITIATION (AS) AND REPLICATION TERMINATION (DS). A) PRINCIPLE OF AS/DS DETECTION BY OKAZAKI-FRAGMENT SEQUENCING. Left: Replication fork directionality (RFD) is calculated by the number of Watson and Crick strands. Right: Theoretical RFD profiles for an initiation/ termination site (for one fixed origin) or zone (multiple inefficient origins). B) VISUAL COMPARISON OF AS AND DS WITH PRE-RC (T-PIC) POSITIONS [chr8: 9033838-9569826]. C) COMPARISON OF AS/DS REGION SIZES. Represented in boxplot: thick line shows the median, the box is the distribution from the first to the third quartile, the whiskers indicate the smallest and largest value without being an outlier. D) PRE-RC (T-PIC) DENSITY WITHIN AS/ DS REGIONS AND THE REMAINING GENOME. Density is represented as events/Mb. AS (initiation zones) are colored in red, DS (termination zones) colored in blue, the remaining genome is represented in grey.

COVERAGE OF PRE-RC COMPONENTS IS ENHANCED AT AS AND DECREASED AT DS

Examining pre-RC positions already procured an impression of pre-RC distribution in relation to AS and DS. Another way of analyzing the data is to calculate the sequencing read coverage at the superimposition of either AS or DS. This has the advantage of being independent of any peak-calling algorithms and obtaining a global impression of the average situation at all sites simultaneously.

Coverage was computed as reads per base within a window of 200 kb around AS or DS centers for each replicate. From this, I calculated the mean coverage of all three replicates for each pre-RC protein. In Figure 4.18, the result of the mean coverage analysis at AS or DS is depicted for A) Orc2, B) Orc3, C) Mcm3, and D) Mcm7. All four pre-RC proteins showed an enhanced coverage at AS, while coverage was depleted from DS. The input is also plotted as a control for intrinsic chromatin composition. As sonication does not fragment chromatin uniformly, regions with higher accessibility might be underrepresented because of increased destruction during sonication, while less accessible regions are slightly overestimated. This is possibly the reason for decreasing input coverage at AS and increasing input coverage at DS. AS have been shown to reside in rather euchromatic environment (Petryk *et al.* 2016), whereas input coverage implies DS being more heterochromatic.

The coverage increase of target pre-RC proteins of 0.5 to 0.8 reads/base seems slight, but accounts for 10% of the mean coverage (assuming a mean coverage of 4.5 reads/base). One possible reason is the large AS/DS size. Looking at an average of all AS ranging from 3.9 kb to 198 kb, also averaged out specific protein binding events. Furthermore, increased coverage at AS had an average region width of ~60 kb and decreased coverage at DS ~120 kb, which corresponds to the mean AS (37.8 kb) and DS sizes (122.2 kb). This observation and the reproducibility of all pre-RC components increasingly covering AS together with a specific coverage depletion at DS convincingly argues for specificity of these results.

In conclusion, both T-PIC-defined pre-RC complex distribution and pre-RC component coverage analysis revealed an elevated pre-RC binding in replication initiation zones (AS), with specific pre-RC depletion in termination zones (DS). When looking at the T-PIC-defined pre-RC complex distribution, pre-RC density was 2.5-fold higher in initiation zones than in termination zones. Analysis of HOMER-defined pre-RC positions by trend also resulted in increased density within AS, but the total number of 329 HOMER-defined pre-RCs was not elevated enough to make a meaningful statement (data not shown).

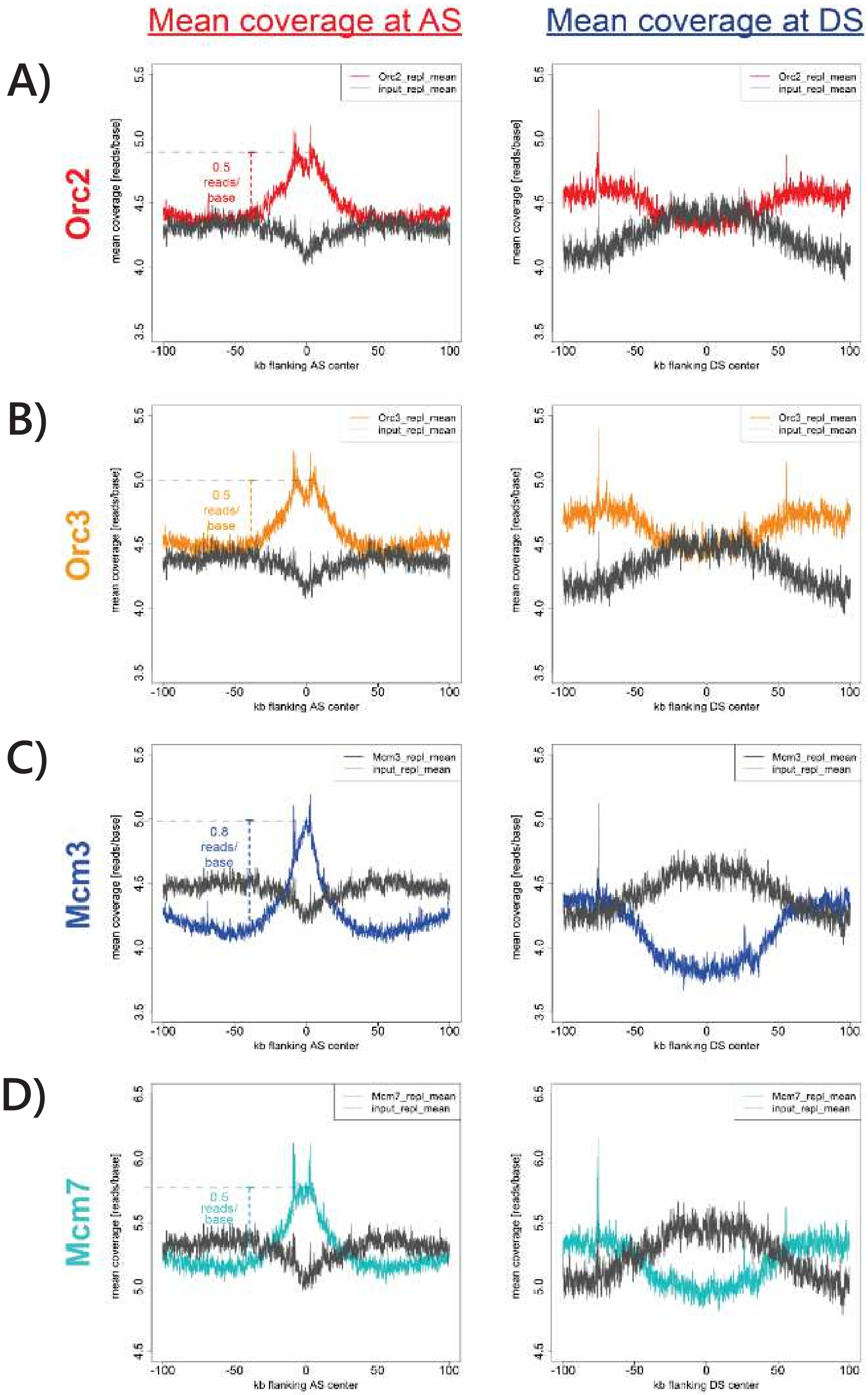

Mean coverage at AS
Mean coverage at DS
A)
Orc2
Orc2_repl_mean
input_repl_mean
0.5 reads/ base
B)
Orc3
Orc3_repl_mean
input_repl_mean
0.5 reads/ base
C)
Mcm3
Mcm3_repl_mean
input_repl_mean
0.8 reads/ base
D)
Mcm7
Mcm7_repl_mean
input_repl_mean
0.5 reads/ base
mean coverage [reads/base]
3.5
4.0
4.5
5.0
5.5
6.0
6.5
-100
-50
0
50
100
kb flanking AS center
kb flanking DS center

FIGURE 4.18: PRE-RC COMPONENT COVERAGE ANALYSIS REVEALED ACCUMULATION AT AS (LEFT) AND DEPLETION AROUND DS (RIGHT). A) MEAN ORC2 COVERAGE. B) MEAN ORC3 COVERAGE. C) MEAN MCM3 COVERAGE. D) MEAN MCM7 COVERAGE. Coverage was calculated as number of reads/base within a 200 kb window around either AS or DS center.

AS are broad zones in which replication preferentially initiates from multiple inefficient origins (Petryk *et al.* 2016). Thus, multiple pre-RC sites are expected within these zones. However, presence of pre-RCs outside of these zones (DS, remaining genome) is not astonishing. These pre-RCs might be part of dormant origins, i.e. origins, that are licensed but not activated, unless replicative stress (replication fork stalling) makes their activation necessary to complete genome replication (Blow, Ge, and Jackson 2011). These observations indicate, that while AS is indeed a zone of pre-RC occurrences and DS clearly is depleted from pre-RCs, origin licensing also efficiently takes place outside of AS.

AS and DS together make up only 15.2% of the human genome, mostly due to low OK-sequencing coverage, which partly impedes confident zone detection. But also in HeLa cells, only 61% of the genome contribute to zones of replication initiation and termination. Petryk *et al.* propose AS as “master initiation zones”, while the rest of the genome displays disperse, stochastically firing origins that cannot be detected by their method. Indeed, this is in agreement with AS representing replication units. Other methods (bubble trap or SNS-seq) detect many active replication events inside and outside AS.

INTENDED IMPROVEMENTS AND FURTHER ANALYSES

This analysis of correlating pre-RC components and replication initiation/termination represents a first impression of relations between origin licensing and activation. Further analyses are anticipated to strengthen these results. First of all, the mean pre-RC density was calculated for all AS/DS. However, this calculation ignores fluctuations. Pre-RC density per single AS/DS needs to be calculated and plotted in a boxplot to be able to draw any final conclusions. If the mean density is influenced by outliers, it will be uncovered. Still, Petryk *et al.* compared HeLa AS to ~13000 Orc1 sites detected in the same cell line by Dellino *et al.* 2013. They found Orc1 peak density being 2.4-times higher in AS than DS, consistent with my complex densities (pre-RC: 2.5-fold higher in AS than DS (Figure 4.17); ORC: 2.4-fold (Appendix Figure 18 B, G1), Mcm2-7: 3.3-fold (Appendix Figure 18 C, G1)).

Furthermore, pre-RC components are not necessarily expected in the middle of AS, as calculated. On the contrary, HeLa Orc1 sites were mostly enriched at AS borders (Petryk *et al.* 2016). Indeed, the M-shaped pre-RC component coverage profiles could result from elevated border distributions (Figure 4.18 A and B). Simple coverage calculations at AS/DS aligned at one border will answer this question. Thereby, one important feature has to be taken into account: Petryk *et al.* found AS often to be flanked by actively transcribed genes. This and the fact the Orc1 sites were often found at AS borders led to the hypothesis, that ORC loads Mcm2-7 helicases at the borders. Delimitation by active transcription possibly directs Mcm2-7 spreading towards AS. When calculating coverage at AS borders, flanking genes need to be taken into account (possibly sorting AS into different classes according to their association with transcription). Two results can be expected: i) indeed, ORC is enriched at AS borders, ii) Mcm3 and Mcm7 are also enriched at the borders but migrate smoothly in the AS direction. Performing the analysis will reveal whether this hypothesis holds true.

CONCLUSION

Biologically, pre-RCs are a prerequisite for active replication. Notably, these results validate my pre-RC ChIP-seq approach and undoubtedly prove that I successfully chipped pre-RC components. While they were nicely associated with AS, the evident depletion at DS was unexpected. Petryk *et al.* did not observe or discuss any similar observations. DS are generally larger than AS (Figure 4.17 C) arguing for replication initiation being more precise than replication termination, which appears to be more random. This observation also indicates that replication fork speeds fluctuate, generating more disperse replication termination sites when forks collide. Pre-RC depletion in DS indicate little origin licensing within sites of replication termination. Termination zones often consist in actively transcribed gene bodies, but have not been further characterized so far. Consequently, several questions arise: Which features define replication termination zones? Is there also a link to heterochromatin? Why are pre-RCs specifically absent from these zones? Is it actually simply the absence of pre-RCs that defines a termination zone as termination zone? And how is active transcription linked to replication termination zones? The last question will be also discussed in the next chapter.

Detection of active replication by any method (OK-seq or SNS-seq) often links DNA replication to active transcription. Conserved AS that are shared between different cell types, are often flanked by actively transcribed genes (Petryk *et al.* 2016). Also SNS are often enriched in genes or promoters (Cayrou *et al.* 2011; Besnard *et al.* 2012; Picard *et al.* 2014; Cayrou *et al.* 2015). Consequently, after correlating pre-RC positions with sites of active replication, I was wondering, whether pre-RC ChIP-seq data also associate with active transcription.

4.3 Stable Origin Licensing is Associated to Regulation of Active Transcription

To be able to correlate pre-RC positions with active transcription, I first needed information about gene expression in Raji cells. Alexander Buschle and Prof. Dr. Wolfgang Hammerschmidt (Research Group EBV Genetics and Vectors, Research Unit Gene Vectors, HelmholtzZentrum München) kindly shared their RNA-seq data obtained from identical Raji cells.

4.3.1 Pre-RC coverage was increased at active transcription start sites

RNA-seq data reflects gene expression within a cell population. The obtained reads from sequencing are mapped against genes and quantified. The experiments were performed in triplicates and the mean quantification was considered to estimate gene expression. With the support of Tobias Straub (head of the Bioinformatics Core Unit, Ludwig-Maximilians-Universität München), actively transcribed genes were arbitrarily defined as more than $e^2 = 7.39$; ($\ln(\text{mean}) > 2$), meaning that more than 7.39 sequencing reads per gene define a gene as actively transcribed. This resulted in 10642 active and 12731 inactive genes. The transcription start sites of these genes were extracted and the coverage of the different pre-RC target proteins was computed within a 4000 bp window around active or inactive TSSs.

As observed in Figure 4.19 (left panel), all targeted members of the pre-RC were clearly enriched at active TSSs. Coverage of Mcm3 (Figure 4.19 C, left panel) was less prominent than for the other target proteins (Figure 4.20: Orc2: A, Orc3: B, and Mcm7: D, left panel), which might originate from the antibody. Pre-RC association was clearly dependent on transcriptional activity, as inactive TSSs were not enriched in any pre-RC component (Figure 4.19, right panel).

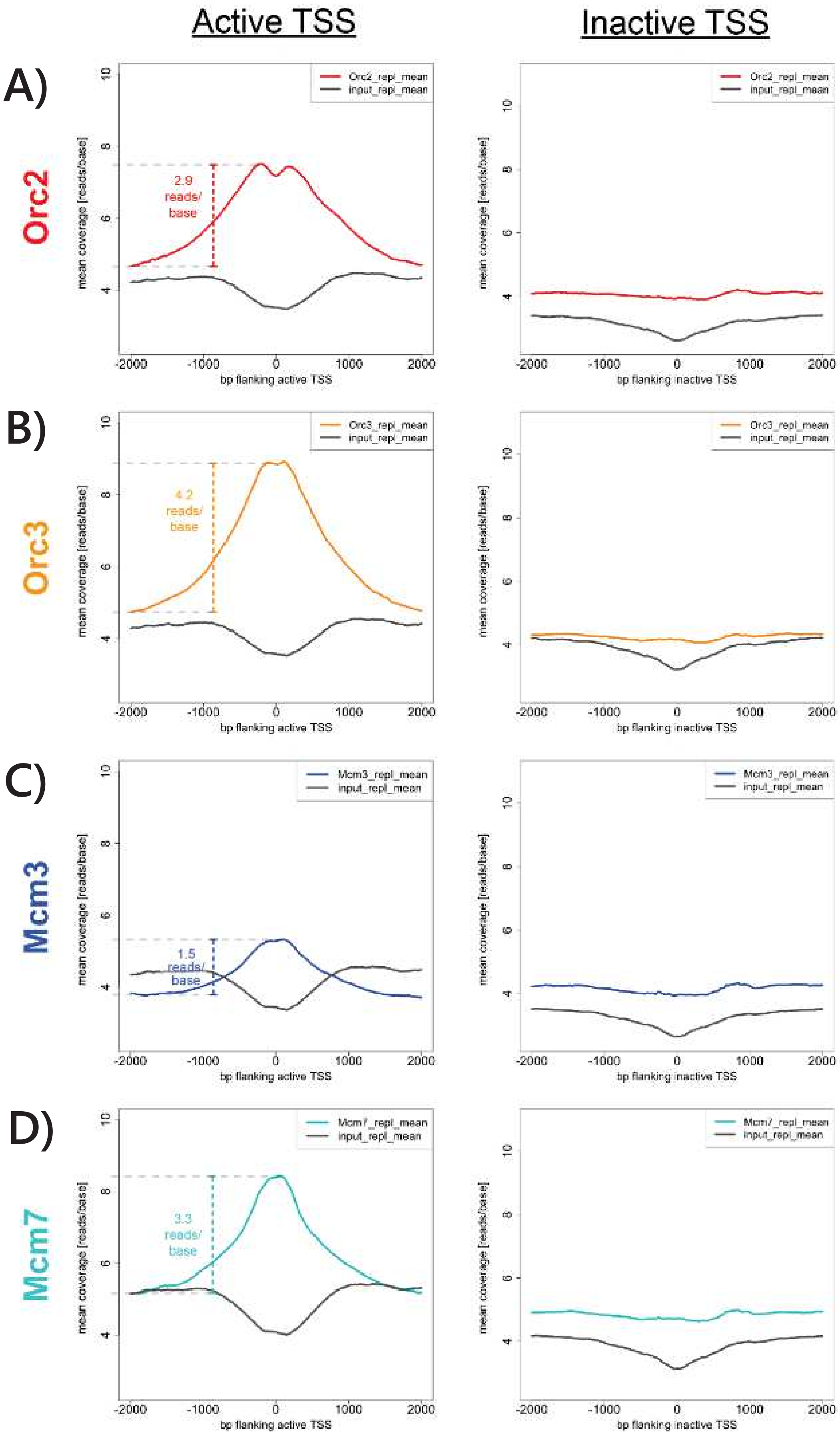

Active TSS
Inactive TSS
A)
Orc2
B)
Orc3
C)
Mcm3
D)
Mcm7
Orc2_repl_mean
input_repl_mean
Orc3_repl_mean
Mcm3_repl_mean
Mcm7_repl_mean
2.9 reads/ base
4.2 reads/ base
1.5 reads/ base
3.3 reads/ base
mean coverage [reads/base]
bp flanking active TSS
bp flanking inactive TSS
-2000
-1000
0
1000
2000
4
6
8
10

FIGURE 4.19: PRE-RC COMPONENT COVERAGE ANALYSIS REVEALED INCREASED COVERAGE AT ACTIVE TSS (LEFT PANEL) AND UNCHANGED COVERAGE AT INACTIVE TSS (RIGHT PANEL). A) MEAN ORC2 COVERAGE. B) MEAN ORC3 COVERAGE. C) MEAN MCM3 COVERAGE. D) MEAN MCM7 COVERAGE. Coverage was calculated as number of reads/base within a 4000 bp window around TSS. Input was plotted as control (grey).

4.3.2 STRONG HOMER-DEFINED ORC, MCM2-7 AND PRE-RC ASSOCIATED WITH ACTIVE TRANSCRIPTION START SITES

To examine the relation of HOMER- or T-PIC-defined ORC, Mcm2-7 and pre-RC to active and inactive TSSs, I calculated the distance of each complex position to the next TSS. This resulted in an estimation of the frequency, in which complexes are found in vicinity or in distance of TSSs. The distances to active TSSs (colored) and inactive TSSs (grey) are represented in the graphs (Figure 4.20: HOMER-defined complexes (left panel); T-PIC-defined complexes (right panel)). As for inactive TSSs, all HOMER- and T-PIC-defined complexes peaked at 100 kb distance. Considering roughly 13000 inactive TSS and a human genome size of 3.1 Mb (hg19), a random distribution would result in an average distance of 238 kb. However, as the genomic organization is not random, but grouped in gene dense euchromatic and gene devoid heterochromatic regions, I propose that ORC, Mcm2-7 and pre-RC distributions around 100 kb from the next inactive TSS originate from TSS-independent distributions. When looking at the T-PIC-defined complexes, their distribution in relation to active TSSs approximate the one of inactive TSSs (Figure 4.20 (right panel)), suggesting that T-PIC-defined pre-RC depend only little on transcriptional activity. Moreover, Mcm2-7 are even more depleted from active TSS, compared to inactive TSS (mean distance from active TSS: 430 kb, mean distance from inactive TSS: 186 kb, p-value = 2.2x10^{-16}).

By contrast, HOMER-defined complex distribution in respect to active TSSs is bipartite (Figure 4.20 (left panel)). A subpopulation of strong HOMER-defined complexes, is located in close proximity (100 – 1000 bp) to active TSSs. This is true for ORC, Mcm2-7 and pre-RC, but also for the single replicates of each ChIP (data not shown). HOMER-defined complexes represent a population of very strong binding sites (Chapter 4.2.1, p. 60). Consequently, the proximity to active TSSs defines at least in part strong pre-RC binding.

Dedicated programs take peak position information and calculate the peak proportions in specific genome features, as e.g. promoters, exons, and introns. Analysis of the genomic distribution of HOMER- and T-PIC-defined pre-RC with the CEAS program (cis-regulatory element annotation system (Shin *et al.* 2009)) confirmed the close association of pre-RCs to proximal promoter regions (see Appendix Figure 11). Thereby, HOMER-defined pre-RCs again showed an increased association to close promoter regions and was also more depleted from intronic and distal intergenic regions compared to T-PIC-defined pre-RCs.

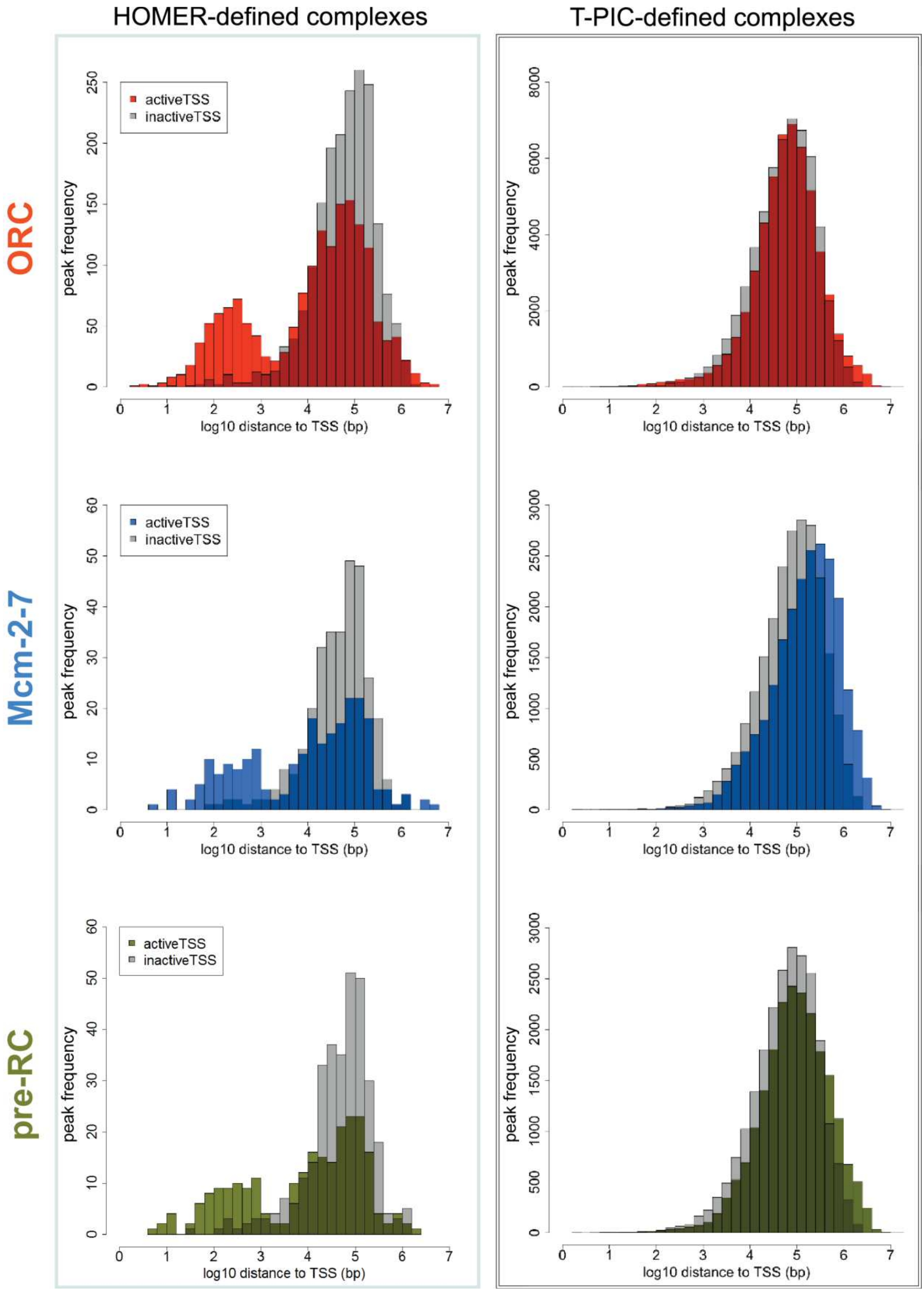


FIGURE 4.20: A SUBPOPULATION OF HOMER-DEFINED ORC, MCM2-7, AND PRE-RC WERE CLOSELY RELATED TO ACTIVE TSS. Left: HOMER-defined complexes, right: T-PIC-defined complexes. The distance of each position was calculated towards the next TSS and is plotted in log10 on the x-axis. Y-axis represents the frequency of complexes within a specific distance.

Nevertheless, T-PIC-defined pre-RCs were also slightly enriched in proximal promoter regions, indicating that active transcription indeed favors pre-RC positions, independent of the peak-calling algorithm.

Promoter activity is amongst others regulated by specific histone modifications. In particular, H3K4me3 is known to recruit chromatin remodeling factors that create open chromatin which leads to transcription factor binding (Kimura 2013). Consequently, most active TSSs are marked by H3K4me3.

4.3.3 PRE-RC COVERAGE WAS ENHANCED AT H3K4ME3 PEAKS

With active TSS being marked by H3K4me3, pre-RC components are also expected to be enriched at H3K4me3 sites. The group of Dr. Jean-Christophe Andrau (Transcription and Epigenomics in Developing T-Cells, IGMM, Montpellier) kindly shared their data of H3K4me3 and H3K36me3 peak positions obtained from ChIP-seq in Raji cells. Indeed, direct comparison of active TSS and H3K4me3 positions revealed that 81% of active TSS coincide with H3K4me3 (Figure 4.21). Still, 52.8% of all H3K4me3 peaks do not directly overlap with TSS. However, further analysis of the genomic distribution of H3K4me3 peaks revealed a close association to all regulatory regions upstream of the TSS (promoter, 5'UTR), as well as presumably first introns and exons (Appendix Figure 12). Computational analysis of the mean coverage of the single pre-RC proteins (Figure 4.22) indicated an enrichment at H3K4me3 peak centers (except for Mcm3 (Figure 4.22 C)). The calculated coverage is more irregular than at TSSs because H3K4me3 peaks of different sizes were centered and read coverage was calculated in a large 20 kb window (mean size = 2.2 kb, range from 250 bp to 98.1 kb, see also Appendix Figure 13).

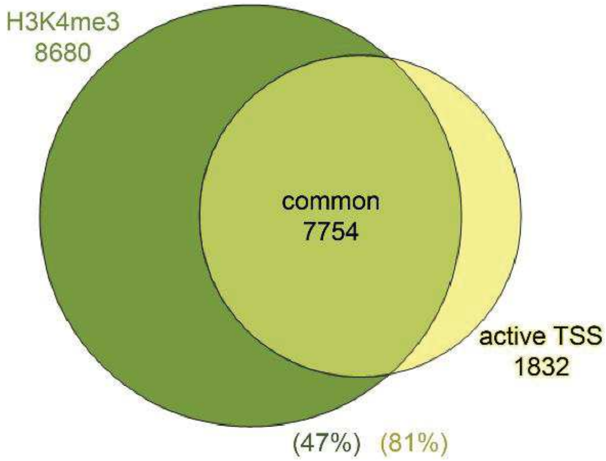


FIGURE 4.21: MOST ACTIVE TSS DIRECTLY OVERLAPPED WITH H3K4ME3. Venn diagram of overlap between TSS and H3K4me3. Overall counts are indicated. The percentage of overlapping proportions are specified below in brackets.

In contrast to pre-RC enrichments at H3K4me3 peaks, no enhanced coverage was observed at H3K36me3 regions (Appendix Figure 15). H3K36me3 is also a histone modification associated to active transcription, but is rather found in actively transcribed gene bodies (Kimura 2013), more precisely in intronic regions (Appendix Figure 14). This finding specifically links replication licensing to the regulation of active transcription, but not to transcriptional activity itself.

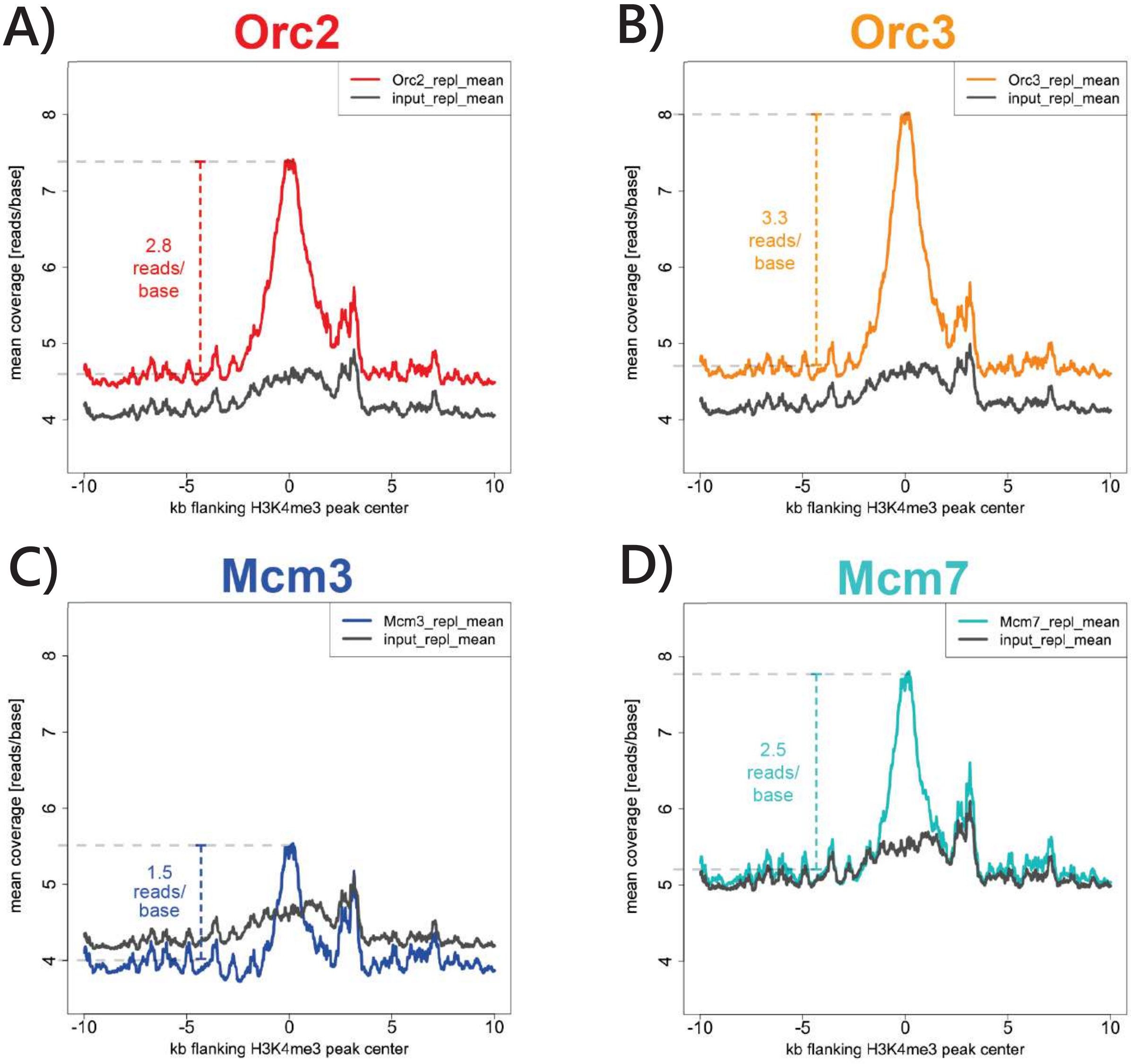


FIGURE 4.22: INCREASED COVERAGE OF TARGET PRE-RC PROTEINS AT H3K4ME3 PEAKS. A) MEAN ORC2 COVERAGE. B) MEAN ORC3 COVERAGE. C) MEAN MCM3 COVERAGE. D) MEAN MCM7 COVERAGE. Coverage was calculated as number of reads/base within a 20 kb window around H3K4me3 peak center. Input was plotted as control (grey).

4.3.4 STRONG HOMER-DEFINED MCM2-7 AND PRE-RC ENTIRELY ASSOCIATE WITH H3K4ME3

Already coverage analyses implied a strong association of pre-RC with H3K4me3. However, it remained to be tested whether this association is confirmed by a co-localization of H3K4me3 and HOMER- and T-PIC-defined ORC, Mcm2-7 and pre-RC positions. Indeed, HOMER-defined Mcm2-7 complexes and pre-RCs nearly completely overlapped with H3K4me3 positions (88%) and ORC at least to 56,3% (Figure 4.23, left panel). In contrast, for T-PIC-defined peaks, ORC positions correlated to a higher degree with H3K4me3 peaks (14.4% of all ORC overlaps with nearly 50% of all H3K4me3 peaks, Figure 4.23, right panel). Mcm2-7 complexes and pre-RCs remained less associated. This result was also confirmed by analyzing the distance to H3K4me3 peak center (Appendix Figure 16, left panel). HOMER-defined Mcm2-7 and pre-RC completely resided in close proximity to the peak centers (100-1000 bp), while a subpopulation of ORC remained distant. The distance of all HOMER- or T-PIC-defined complexes to H3K36me3 remained randomly distributed.

CONCORDANCE WITH PREVIOUS STUDIES AND FURTHER ANALYSES

The regulation of transcriptional activity clearly impacts on origin licensing efficiencies. The comparison of HOMER- and T-PIC-defined complexes argues for the variability of favorable origin licensing features. Strong licensing depends entirely on active transcriptional regulation (association to TSS and H3K4me3) and thus affects the entire HOMER-defined pre-RC population. Regarding T-PIC-defined ORC, Mcm2-7 and pre-RC, licensing positions regulated by active transcription only built a sub-population of all T-PIC peaks and further features account for the other T-PIC-defined complexes. However, these features confer less localized binding of pre-RC components.

Interestingly, ORC is the only HOMER-defined complex that did not entirely overlap with H3K4me3 peaks. This hypothesis is consistent with the fact that ORC also has other functions than regulating pre-RC formation, for instance in heterochromatin regulation (Giri *et al.* 2015; Giri and Prasanth 2015). This can be tested by extracting the ORC subpopulation not associating with H3K4me3 and comparing these ORC positions with known heterochromatin regions.

The association of active DNA replication and transcription is already widely proven (Martin *et al.* 2011; Besnard *et al.* 2012; Cayrou *et al.* 2015; Smith *et al.* 2016; Kylie *et al.* 2016). Furthermore, also both ChIP-seq studies targeting a component of ORC (Orc1: Dellino *et al.* 2013, Orc2: Miotto, Ji, and Struhl 2016), stated a close connection to transcriptional activity. Dellino *et al.* found 71% of all detected Orc1 sites at active TSS.

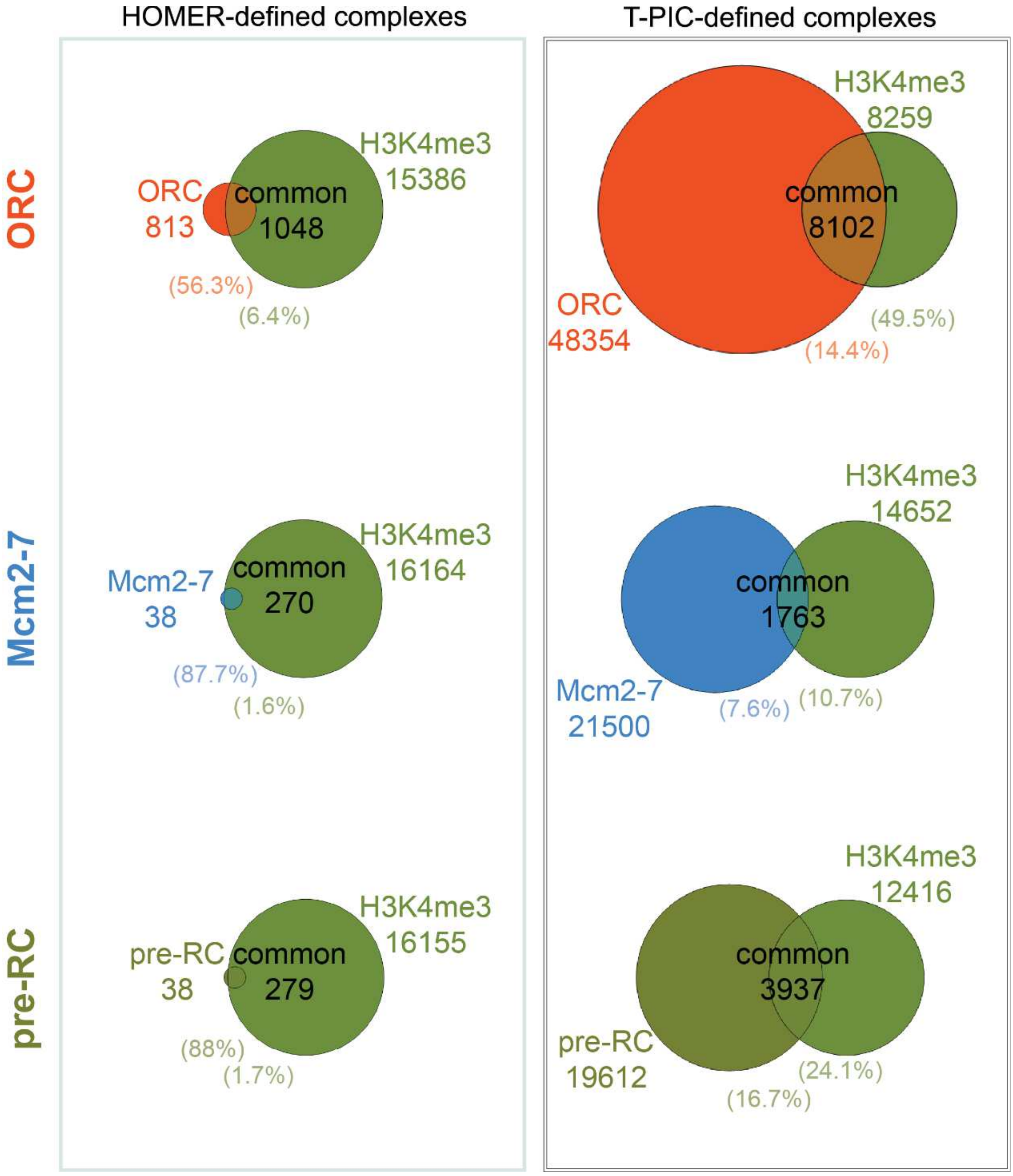


FIGURE 4.23: HOMER-DEFINED COMPLEXES NEARLY COMPLETELY OVERLAP WITH H3K4ME3 PEAKS. Left panel: HOMER-defined complexes, right panel: T-PIC-defined complexes. Venn diagram of the overlap between ORC, Mcm2-7, and pre-RC with H3K4me3 peaks. Overall counts are indicated. The percentages of overlapping proportions are specified below in brackets.

Additionally, the authors linked Orc1 sites with high transcription levels to early replication, while Orc1 sites with low or undetectable expression levels replicated late in S-phase. Thus, their study connects Orc1 sites and transcription with replication timing.

Miotto *et al.* claim only a moderate association of Orc2 sites and active transcription (Pearson correlation 0.33, Miotto, Ji, and Struhl 2016). Considering ~52000 detected Orc2 sites and roughly 20000 human genes, of which maybe half are transcriptionally active, only $\frac{1}{5}$ of all Orc2 peaks *could* actually correlate with active transcription and a correlation coefficient of 0.33 should be judged as good. Re-analyzing their data and sorting Orc2 peaks for strong and weak enrichments might also connect strong Orc2 binding to active TSS. I intend to directly correlate their Orc2 peak positions with my HOMER- and T-PIC defined ORC, Mcm2-7, and pre-RC, to get an idea of concordance of the experiments. The authors claim that DNA accessibility is the main determinant of Orc2 positions. Consequently, they found Orc2 in promoters and regions enriched for active chromatin marks (H3K27ac, H3K4me1/2/3), consistent with my own results.

Very recently, Smith *et al.* compared active replication in different human cell types and differentiation states and found mostly conserved replication origins associated with H3K4me3 (Smith *et al.* 2016). Also Cayrou *et al.* linked active replication to histone modifications favoring open chromatin (H3K4me, H3K9Ac, Cayrou *et al.* 2015). These findings suggest H3K4me3 being a histone modification that contributes to an accessible chromatin environment for strong pre-RC binding and thus efficient replication initiation.

The general question remains whether transcriptional activity positively or negatively influences origin licensing and activation. While it is evident that active transcription creates a chromatin environment also accessible for replication factors, multiple studies mutually exclude high transcription rates and efficient replication (Nieduszynski, Blow, and Donaldson 2005; Mori and Shirahige 2007; Lõoke *et al.* 2010; Martin *et al.* 2011). My results imply, that actively transcribed genes are indeed devoid of pre-RC components (absent pre-RC coverage at H3K36me3 peaks, Appendix Figure 15), while pre-RCs are mostly found at active TSS (Figure 4.19) and promoter regions harboring active histone modification marks (H3K4me3, Figure 4.22). It has been reported that replication initiation events themselves are absent from TSS, but enriched in adjacent sequences (Martin *et al.* 2011). Petryk *et al.* found early replicating initiation zones predominantly flanked by actively transcribed genes, while many termination zones overlap active genes. Assuming that ORC is mostly found at initiation zone borders (as shown for Orc1, Dellino *et al.* 2013; Petryk *et al.* 2016), and Mcm2-7 helicases move from their initial loading site, one could hypothesize that transcription and RNA polymerase II action promote Mcm2-7 removal from actively transcribed genes.

RNA polymerase dependent Mcm2-7 double-hexamer sliding has indeed been shown by Gros *et al.* 2015 in budding yeast. The authors found that RNA polymerase II can push Mcm2-7 double-hexamers along the DNA, without them losing initiation potential. Also in *Leishmania major*, it has recently been demonstrated that most replication initiation sites are found at sites of RNA polymerase II stalling or transcriptional termination, arguing for co-migration of Mcm2-7 helicases and the transcriptional machinery (Lombraña *et al.* 2016). Furthermore, Powell *et al.* demonstrated by ChIP-seq studies in *Drosophila melanogaster*, that Mcm2-7 helicases are evenly distributed throughout the genome in G1/S transition, but absent in actively transcribed genes (Powell *et al.* 2015). Thereby, an active transcription process is required for Mcm2-7 displacement from gene bodies. The authors propose Mcm2-7 residing in transcribed genes being displaced by the passage of RNA polymerase II.

Absence of Mcm2-7 helicases at actively transcribed genes would explain active genes being favored zones of replication termination (Petryk *et al.* 2016). This model would situate ORC localizing at accessible regions, like TSS, and thus initiation zone boundaries. ORC loads Mcm2-7 helicases, which are subsequently translocated by active transcription into gene-adjacent regions, defining replication initiation zones detected by Okazaki-fragment sequencing. This RNA polymerase II-dependent translocation mechanism would also avoid head-to-head collisions of replication fork and the transcriptional machinery, preventing sources of replication fork stalling or DNA damage. Moreover, G4 structures might block Mcm2-7 helicase sliding, possibly explaining why G4 structures are found enriched at active replication origins (Besnard *et al.* 2012). The mechanism of Mcm2-7 dislocation would include displacement or reassembly of nucleosomes and other DNA-binding proteins, which seems laborious for the cell and needs further investigation.

Conclusion

Pre-RC or replication origin association to the regulation of active transcription is generally accompanied by early replication timing. My results showing strong pre-RC binding at active TSS provoke the speculation whether active TSS are early replicated *because* of strong pre-RC binding. Miotto *et al.* used mathematical modeling to correctly predict replication timing dependent on ORC densities (Miotto, Ji, and Struhl 2016). As higher ORC densities presumably also result in more Mcm2-7 helicases, the probability of origin firing events increases, resulting in early replication. Replication timing domains are concordant with topologically associating domains (TADs), whose borders also might limit the Mcm2-7 helicase sliding to restricted domains.

Consequently, active chromatin environment positively influences ORC binding probabilities, which leads to more Mcm2-7 helicase loading. These helicases are translocated from their original binding sites by transcription machineries, which is the reason for replication initiation not taking place within genes, but adjacent to them. The more accessible chromatin is, the higher ORC/Mcm2-7 densities may become, the higher the probability for licensed origins to fire, the earlier replication origins are activated.

4.4 MCM2-7 HELICASES APPROXIMATE ORC IN S/G2

Origin licensing can only happen in late mitosis/ early G1, when CDK and DDK activities are low. After ORC binding to DNA, Cdc6 and Cdt1 are required for further Mcm2-7 double-hexamer loading. Once a replication origin is activated, the helicases move as part of the replication forks, until replication termination and helicase dissociation.

I chose the S/G2 time-point during the cell cycle as a control for my ChIP experiments, assuming that replication complexes either dissociate because of their replication activity, or passively, when replication forks move through inactive licensing complexes, thereby removing them from DNA (Kuipers *et al.* 2011). As a consequence, and as observed at *oriP*, I expected less Mcm2-7 binding to chromatin sites, resulting in less peak-algorithm defined complexes.

4.4.1 COMPLEX NUMBER AND SIZES DO NOT CHANGE IN S/G2

S/G2 pre-RC-ChIPs were performed in parallel to the G1 ChIPs and sequenced comparably. I analyzed the data as described previously and performed HOMER and T-PIC peak-calling, starting with comparing the number and positions of detected HOMER- or T-PIC-defined complexes in G1 and S/G2.

For both HOMER-defined complexes, as well as T-PIC-defined complexes, neither the peak number nor the peak sizes changed considerably (Figure 4.24, for EBV genome, see Appendix Figure 17 A and B). Opposing to what was expected, there were even more HOMER-defined strong Mcm2-7 complexes detected in S/G2, which consequently resulted in slightly more defined pre-RCs (Figure 4.24 A). As for T-PIC-defined complexes, there was a minor reduction in the number of defined complexes observed (Figure 4.24 B). Comparing the positions of the defined-complexes between G1 and S/G2 on the EBV genome revealed a strong conservation (Appendix Figure 17 C). Also for the human genome, the majority of the HOMER-defined complexes in G1 are conserved in S/G2 (> 60%, Figure 4.25, left panel). Because more HOMER-defined Mcm2-7 positions were detected in S/G2, many new pre-RC sites were also determined.

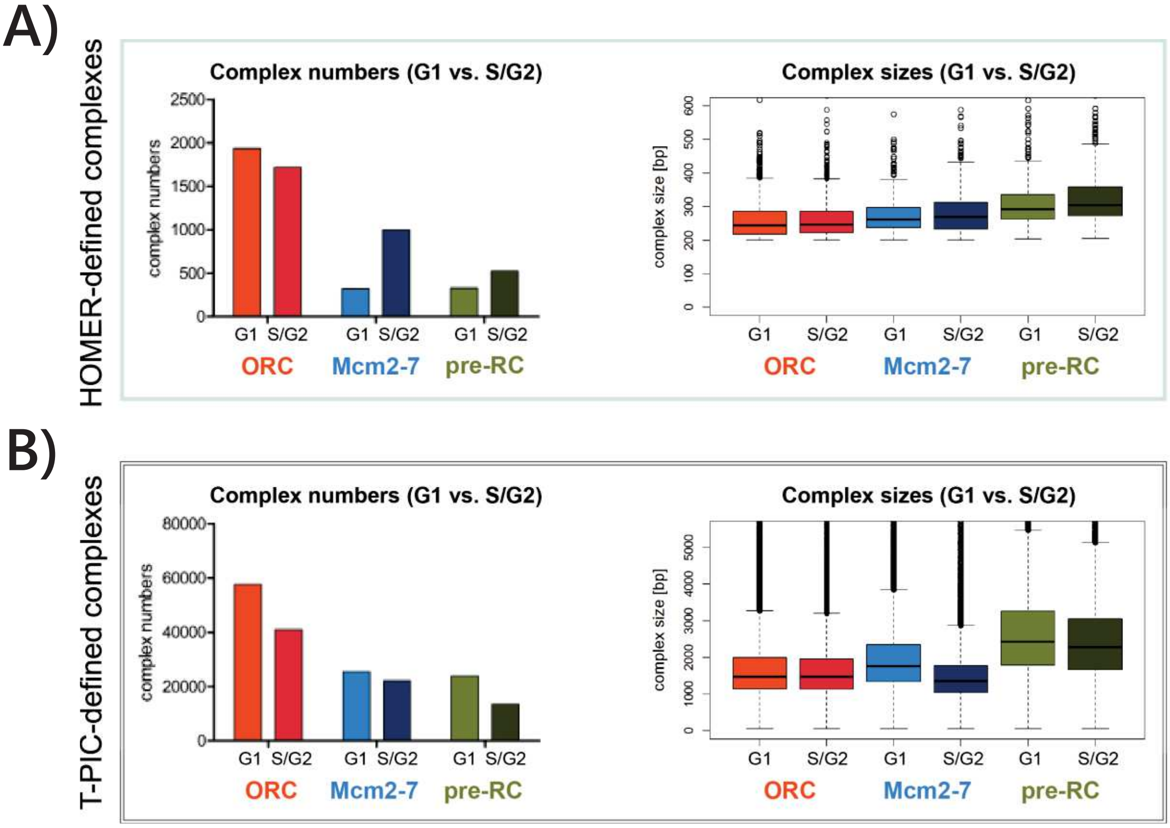


FIGURE 4.24: DEFINED COMPLEXES DO NOT DIFFER IN NUMBER OR SIZE WHEN COMPARING G1 VS. S/G2. A) HOMER-DEFINED COMPLEXES. B) T-PIC-DEFINED COMPLEXES. Left: Number of defined ORC, Mcm2-7 and pre-RC was plotted as indicated in a bar chart in G1 vs. S/G2. Right: Complex sizes were potted as boxplot in G1 vs. S/G2. Thick line shows the median, the box is the distribution from the first to the third quartile, the whiskers indicate the smallest and largest value without being an outlier. Outliers represented by dots.

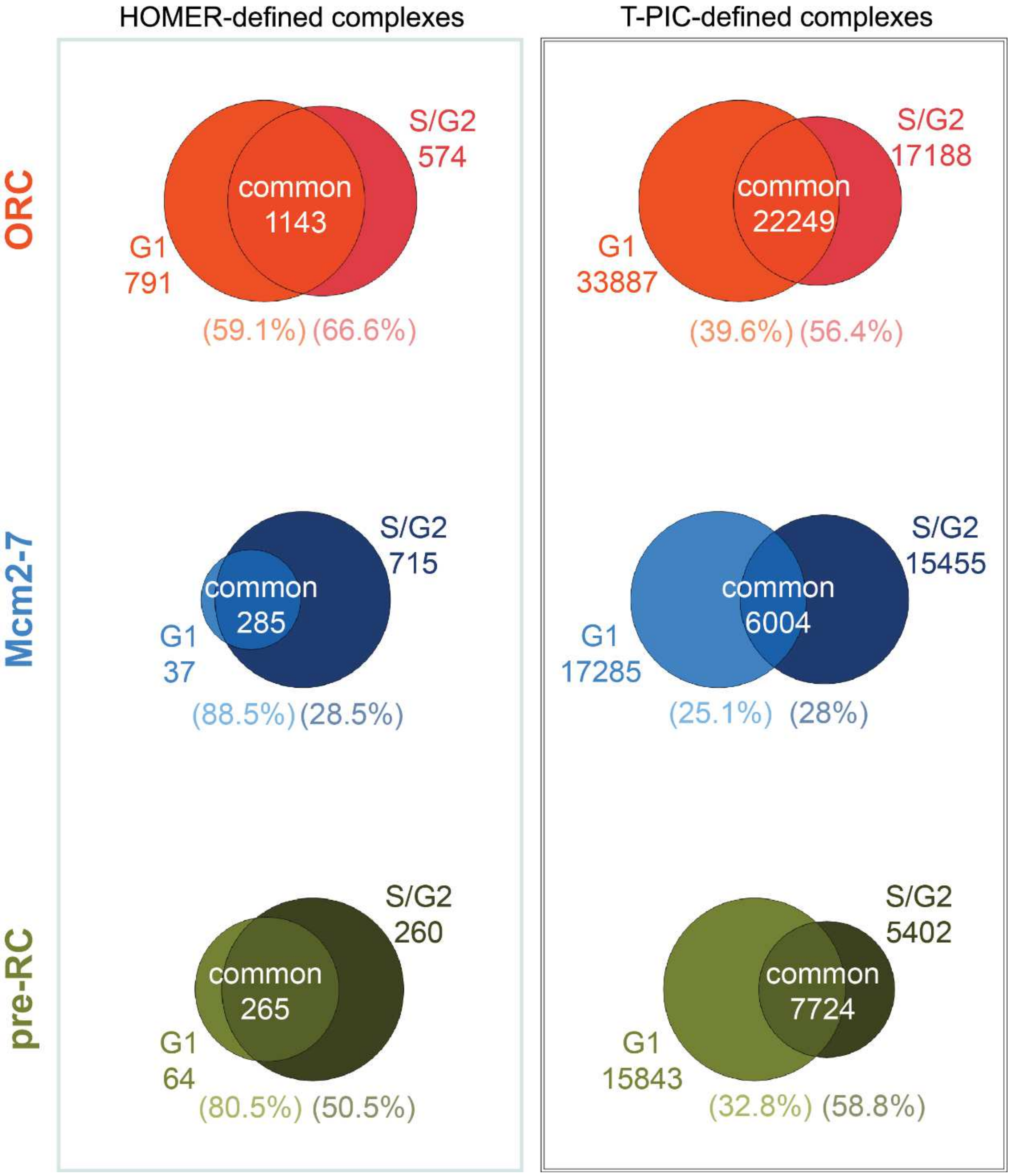


FIGURE 4.25: COMPARISON OF HOMER- AND T-PIC-DEFINED COMPLEX POSITIONS IN G1 VS. S/G2. Left: HOMER-defined complexes, right: T-PIC-defined complexes. Venn diagram of overlap between G1-defined ORC, Mcm2-7, and pre-RC with the same complexes defined in S/G2. Overall counts are indicated. The percentage of overlapping proportions are specified below in brackets.

Concerning T-PIC-defined complexes, the situation was less pronounced. While there was a considerable conservation of ORC and pre-RC positions observed (>55% of S/G2 complexes overlap with G1 positions, Figure 4.25, right panel), this was not the case for Mcm2-7 (< 30% overlap of S/G2 positions with G1). The reduction of the number T-PIC-defined complexes in S/G2 can be explained by a decrease in *global* protein binding. This reduced binding resulted in a more dispersed signal that did not meet the criteria for significant peak detection. However, observed reduction of defined T-PIC complexes was relatively unimportant, as a significant number of ORC, Mcm2-7, and pre-RCs were still detected in S/G2, indicating that a considerable amount of pre-RC proteins were still bound to DNA at this cell cycle stage. Moreover, I detected an increased amount of HOMER-defined Mcm2-7 complexes and pre-RCs, arguing for an increased strong binding of Mcm2-7 proteins at defined positions in S/G2.

I further analyzed the association of S/G2 ORC, Mcm2-7 and pre-RC with replication, transcription and active histone marks. In G1, pre-RC coverage was detected at replication initiation zones and specifically depleted from replication termination zones. In S/G2, after passage of the replication forks, I would expect a flat signal for both replication zones.

4.4.2 MCM2-7 COVERAGE AT AS AND DEPLETION FROM DS ARE LESS PRONOUNCED

Coverage analysis is a measure of global changes in binding of the single pre-RC components to certain features. A direct comparison of the maximum read coverage at AS is shown in Table 4.2, p. 89. While pre-RC coverage at AS decreased in S/G2, especially for Mcm2-7 components (for Mcm3: from 0.8 reads/base in G1 to 0.3 reads/base in S/G2, for Mcm7: from 0.5 reads/base in G1 to 0.3 reads/base in S/G2), coverage at DS seems less depleted (Figure 4.26). Also the quantification of T-PIC-defined ORC, Mcm2-7 and pre-RC densities revealed no major change of ORC densities at AS or DS, but a decrease in the remaining genomic regions (Appendix Figure 18 B). Mcm2-7 densities however showed an increase at DS, while AS and the remaining genome remain unchanged (Appendix Figure 18 C). These density shifts of ORC and Mcm2-7 resulted in pre-RC density reductions mostly at the remaining genome (Appendix Figure 18 A). These observations indicate T-PIC-defined ORC positions to only marginally change within AS or DS, but that the reduction of ORC numbers in S/G2 essentially concerned the remaining genome. T-PIC-defined Mcm2-7 complex densities and Mcm2-7 coverage rather indicate that Mcm2-7 proteins were less present at AS and less absent from DS in S/G2.

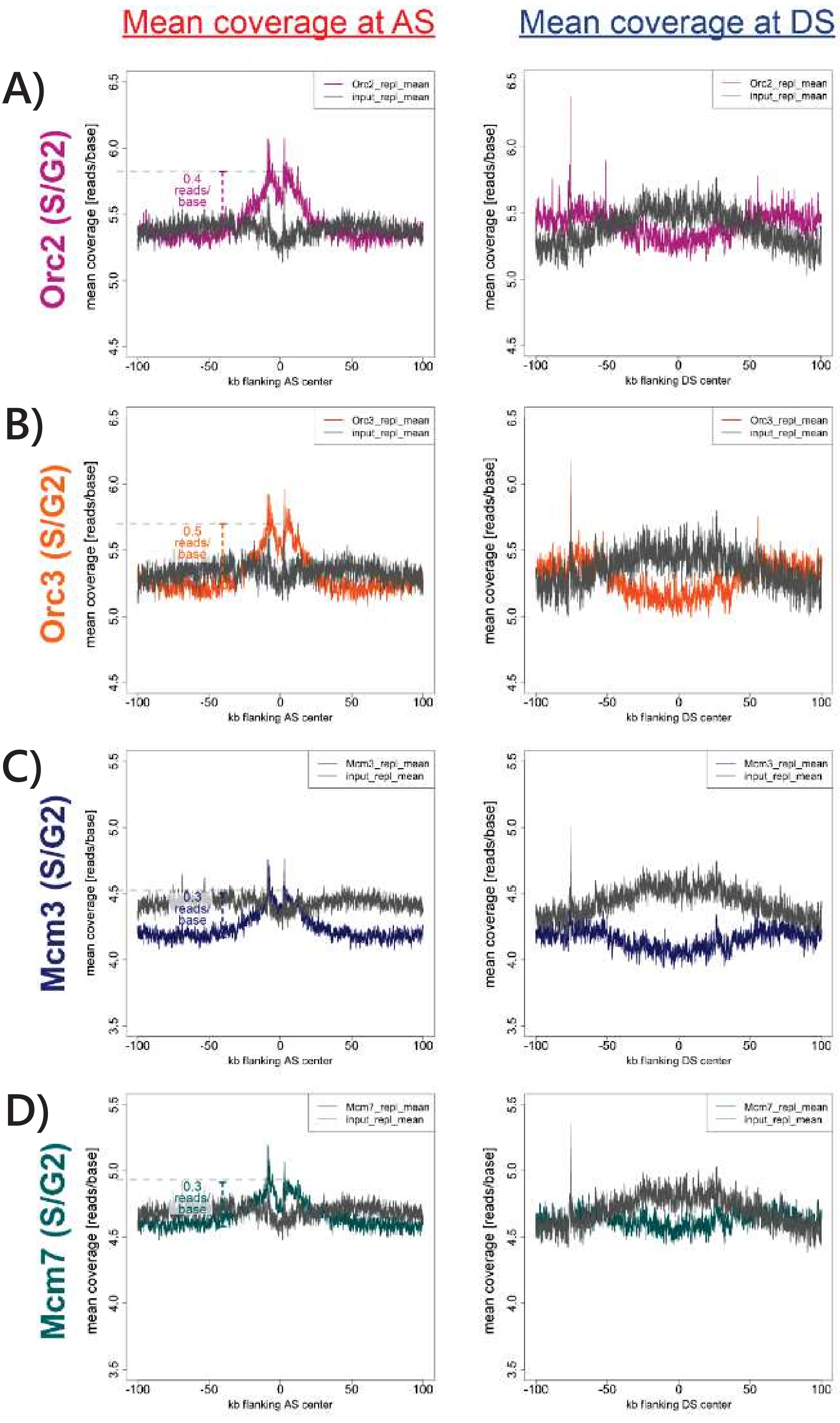

Mean coverage at AS
Mean coverage at DS
A)
Orc2 (S/G2)
B)
Orc3 (S/G2)
C)
Mcm3 (S/G2)
D)
Mcm7 (S/G2)
mean coverage [reads/base]
kb flanking AS center
kb flanking DS center
Orc2_repl_mean
input_repl_mean
Orc3_repl_mean
Mcm3_repl_mean
Mcm7_repl_mean
0.4 reads/ base
0.5 reads/ base
0.3 reads/ base
0.3 reads/ base
-100
-50
0
50
100
3.5
4.0
4.5
5.0
5.5
6.0
6.5

FIGURE 4.26: PRE-RC COMPONENT COVERAGE (S/G2) AT AS (LEFT PANEL) AND DS (RIGHT PANEL). A) MEAN ORC2 S/G2 COVERAGE. B) MEAN ORC3 S/G2 COVERAGE. C) MEAN MCM3 S/G2 COVERAGE. D) MEAN MCM7 S/G2 COVERAGE. Coverage was calculated as number of reads/base within a 200 kb window around AS/ DS region center. Input was plotted as control (grey).

It has already been shown in chromatin binding experiments based on cell cycle fractionation by centrifugal elutriation, that ORC still binds chromatin in S/G2, while Mcm3 and Mcm7 binding generally decreases (Ritzi *et al.* 2003). Indeed, while Orc2 and Orc3 coverage profiles are very similar in G1 and S/G2 (compare Figure 4.18 and Figure 4.26), Mcm3 and Mcm7 profiles assimilate Orc2/3 profiles, which involves decreased coverage at AS and less depletion from DS.

Pre-RC ChIP-seq experiments in *Drosophila* cells showed that minimal Mcm2-7 double-hexamer loading can occur in absence of cyclin E (main regulator of S-phase entry), while maximal loading coincides with G1/S transition. Minimal loading strictly depends on ORC, while maximal Mcm2-7 helicase loading during G1/S coincides less with ORC positions (Powell *et al.* 2015). It is possible that my pre-RC coverage profiles in S/G2 reflect minimal Mcm2-7 loading, strictly dependent on ORC, which might be the reason for both ORC and Mcm2-7 profiles becoming similar.

ORC generally binds in accessible regions, the reason for ORC being mostly detected at TSS or H3K4me3 enriched regions (chapter 4.3, p. 72ff). If minimal Mcm2-7 double-hexamer loading in S/G2 strictly depends on ORC, I expect an increased Mcm2-7 coverage at these accessible regions.

4.4.3 PRE-RC COVERAGE AT TSS AND H3K4ME3 WAS INCREASED IN S/G2

When calculating the coverage of pre-RC subcomponents at active TSS, their association was generally increased (Table 4.2, Orc2 coverage from 2.9 reads/base in G1 to 3.5 reads/base in S/G2, Orc3 coverage from 4.2 reads/base in G1 to 5.1 reads/base in S/G2, Mcm3 coverage from 1.5 reads/base in G1 to 2.8 reads/base in G2, Mcm7 from 3.3 reads/base in G1 to 4.4 reads/base in G2; compare Figure 4.19 (G1) with Figure 4.27 (S/G2)). This enhancement was more prominent for Mcm3 and Mcm7 (> 1 read/base). Enhanced Mcm2-7 binding was also detected by HOMER peak-calling. Strong HOMER peaks were shown to preferentially locate close to active transcription sites (Figure 4.20, left panel; Figure 4.23, left panel; Appendix Figure 11 A; Appendix Figure 16, left panel) consistent with these results.

H3K4me3 peaks were also more covered by pre-RC components in S/G2 (Table 4.2, Orc2 coverage from 2.8 reads/base in G1 to 3.8 reads/base in S/G2, Orc3 coverage from 3.3 reads/base in G1 to 4.3 reads/base in S/G2, Mcm3 coverage from 1.5 reads/base in G1 to 2.5 reads/base in S/G2, Mcm7 from 2.5 reads/base in G1 to 3.2 reads/base in S/G2; compare Figure 4.22 (G1) with Appendix Figure 19 (S/G2)).

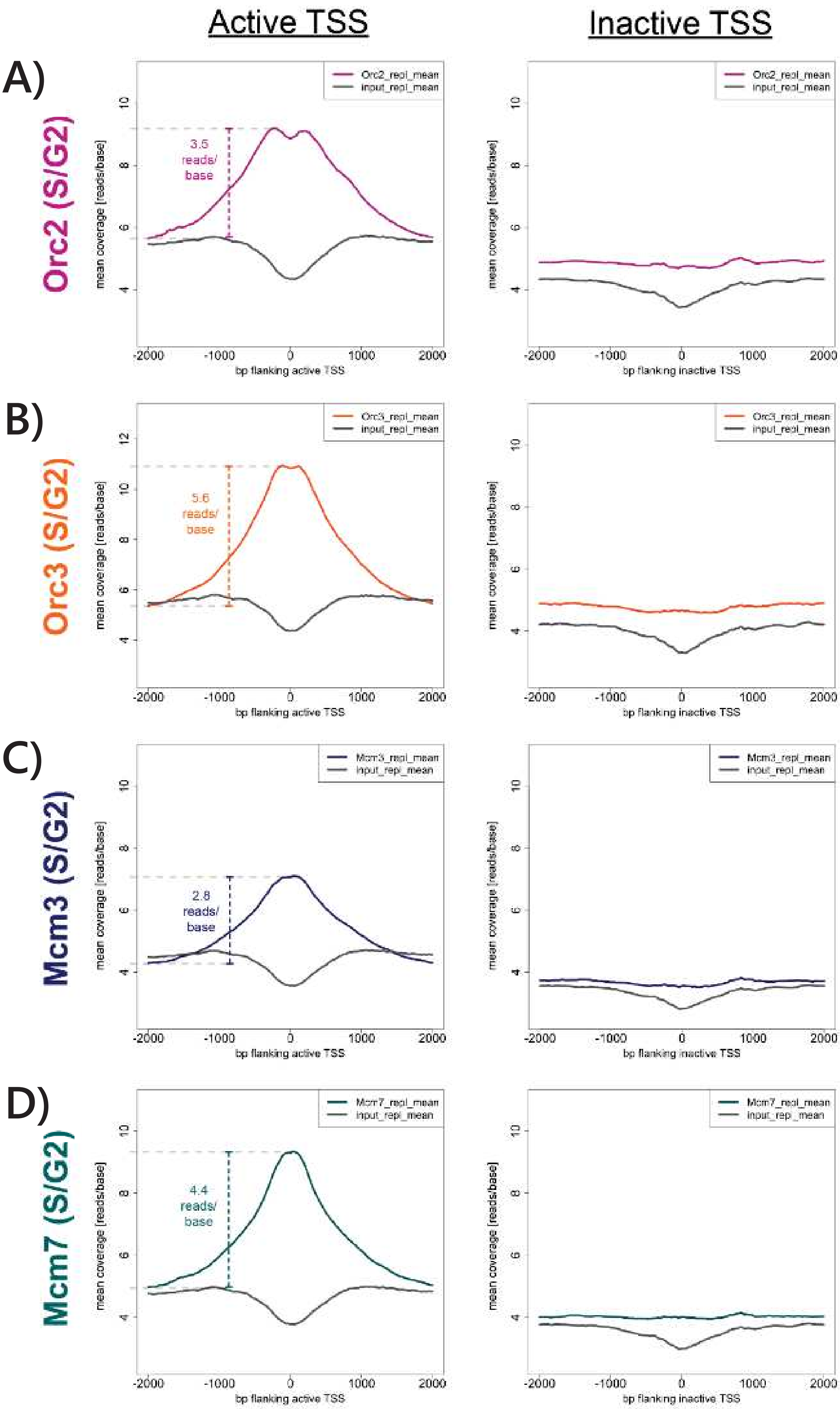

Active TSS
Inactive TSS
A)
Orc2 (S/G2)
Orc2_repl_mean
input_repl_mean
3.5
reads/
base
mean coverage [reads/base]
4
6
8
10
-2000
-1000
0
1000
2000
bp flanking active TSS
bp flanking inactive TSS
B)
Orc3 (S/G2)
Orc3_repl_mean
input_repl_mean
5.6
reads/
base
12
C)
Mcm3 (S/G2)
Mcm3_repl_mean
input_repl_mean
2.8
reads/
base
D)
Mcm7 (S/G2)
Mcm7_repl_mean
input_repl_mean
4.4
reads/
base

FIGURE 4.27: PRE-RC PROTEIN COVERAGE INCREASED AT ACTIVE TSS IN S/G2. A) MEAN ORC2 S/G2 COVERAGE. B) MEAN ORC3 S/G2 COVERAGE. C) MEAN MCM3 S/G2 COVERAGE. D) MEAN MCM7 S/G2 COVERAGE. Coverage was calculated as number of reads/base within a 4000 bp window around TSS. Input was plotted as control (grey). Active TSS (left) inactive TSS (right).

TABLE 4.2: DIRECT COMPARISON OF MAXIMUM MEAN READ COVERAGE AT REGIONS OF INTEREST IN G1 VS. S/G2. Mean read coverage is specified in reads/base.

		G1	S/G2	Difference
Initiation zones (AS)	*Orc2*	0.5	0.4	- 0.1
	Orc3	0.5	0.5	0
	Mcm3	0.8	0.3	- 0.5
	Mcm7	0.5	0.3	- 0.2
TSS	*Orc2*	2.9	3.5	+ 0.6
	Orc3	4.2	5.6	+ 1.4
	Mcm3	1.5	2.8	+ 1.3
	Mcm7	3.3	4.4	+ 1.1
H3K4me3	*Orc2*	2.8	3.8	+ 1.0
	Orc3	3.3	4.3	+ 1.0
	Mcm3	1.5	2.5	+ 1.0
	Mcm7	2.5	3.2	+ 0.7

When directly comparing HOMER- and T-PIC-defined pre-RC positions overlapping with H3K4me3 peaks, it became evident that the S/G2 HOMER-defined pre-RCs still mainly overlap with H3K4me3 (Appendix Figure 20, left panel; for ORC and Mcm2-7, see Appendix Figure 21). Thereby, also a higher proportion of T-PIC-defined pre-RCs associated with H3K4me3 in S/G2 (Appendix Figure 20, right panel).

These observations lead to the conclusion that, while HOMER-defined complexes are still related to active transcription in S/G2, T-PIC-defined complexes seem also to locate more towards actively transcribed regions in S/G2. This hypothesis was also confirmed by analyzing the genomic distribution of HOMER- and T-PIC defined pre-RCs (Appendix Figure 22).

These results indicate that the distribution of Mcm2-7 complexes changed in S/G2. The enrichment in active replication sites decreased as well as the specific depletion from replication termination sites was less detectable. Moreover, the association of Mcm2-7 to active transcription notably increased. One possible explanation for these observations is Mcm2-7 being loaded on DNA without sliding away from their loading site. This possibly reflects an early time point of Mcm2-7 loading and would result a higher similarity between ORC and Mcm2-7 helicases in S/G2.

4.4.4 Mcm2-7 helicases approximate ORC positions in S/G2

To evaluate the relation between ORC and Mcm2-7 peak positions, I calculated the Jaccard index as explained in Figure 4.9 A and represented similarities in a heatmap (Figure 4.28). For both HOMER- and T-PIC-defined complexes, highest similarities were obtained when comparing each complex between G1 and S/G2 (highlighted as red boxes in Figure 4.28), consistent with complex positions being conserved for up to 80% (as shown in Figure 4.25). Only T-PIC defined Mcm2-7 showed little similarity between G1 and S/G2, as expected from previous results. However, similarities between ORC (S/G2) and Mcm2-7 (S/G2) increased (green boxes in Figure 4.28) compared to ORC (G1) and Mcm2-7 (G1). These results confirm Mcm2-7 positions approximating ORC positions in S/G2.

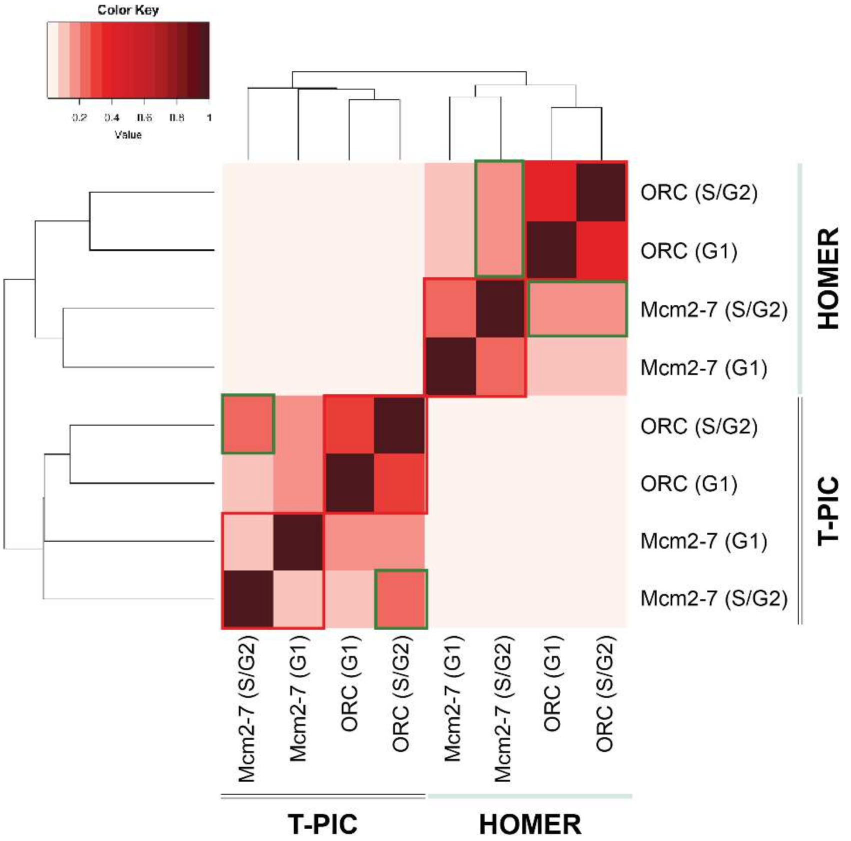


FIGURE 4.28: MCM2-7 APPROXIMATED ORC IN S/G2. T-PIC- and HOMER-defined ORC and Mcm2-7 complex similarities represented as heatmap of Jaccard indices (G1 vs. S/G2). Dark red indicates high similarities while lighter red marks differences. Each complex generally tends to cluster together when comparing G1 vs. S/G2 (marked by red boxes). Mcm2-7 similarity to ORC increases in S/G2 (green boxes).

Possible reasons for premature Mcm2-7 binding and anticipated validation

During my analyses it became evident, that Mcm2-7 peak-calling does not necessarily lead to meaningful biologically relevant results. Due to the discussed Mcm2-7 sliding process, Mcm2-7 complexes may become undetectable by peak-calling algorithms. HOMER detects only accumulations of Mcm2-7, probably reflecting their initial loading sites.

Apparently, Mcm2-7 helicases are more dynamic in G1 (very low number (322) of strong HOMER-defined Mcm2-7), while they seem to be more localized in S/G2 (3x more HOMER-defined Mcm2-7 (1002), Figure 4.24 A).

Thereby, most HOMER-defined Mcm2-7 from G1 are conserved in S/G2 (88.5%, Figure 4.25, left panel). In contrast, T-PIC is more sensible for detecting protein enrichments. It obviously detects HOMER-defined peaks (Figure 4.13, Appendix Figure 6), but also identifies regions of lower intensities, presumably sites of Mcm2-7 helicase spreading. However, this sensible detection can explain low reproducibility of detected peak positions, as stochastic Mcm2-7 helicase sliding is not necessarily always detected at similar sites.

Originally, I planned to use ChIP-seq data from the S/G2 cell cycle phase as negative control for my experiments. Interestingly, my analyses in S/G2 provided insights in previously unknown Mcm2-7 DNA binding processes happening after origin firing, and it seems that pre-RC proteins are not completely dissociated from DNA in post-replicative cell cycle stages. While association of Mcm2-7 to replication units were reduced, ORC and Mcm2-7 were preferentially found at regions of active transcription. Thereby, Mcm2-7 showed a higher similarity to ORC positions in S/G2, than in G1. Consequently, these results suggest an early origin licensing state, with licensing first taking place at accessible regions and Mcm2-7 helicases still localizing at their loading sites.

Origin licensing is generally prevented in G2. Major licensing block is the inhibition of the Mcm2-7 chaperone Cdt1 (Ballabeni *et al.* 2004). Cdk2 (highest expression in G2) phosphorylates Cdt1, thereby targeting it for degradation (Blow and Dutta 2005). Furthermore, Geminin inactivates Cdt1 during G2 and mitosis. Consequently, Mcm2-7 loading can only take place in late mitosis/early G1.

The observation of Mcm2-7 loading in S/G2 might be explained by an elevated contaminating mitotic cell population. Thus, I need to more precisely define the exact cell cycle stage of my S/G2 population. Although centrifugal elutriation is routinely used in my laboratory and has been extensively characterized (Ritzi *et al.* 2003; Papior 2010), FACS analysis (Figure 4.2) is not sufficient to confirm the precise cell cycle stage. Expression of Cdt1 itself, as well as cell cycle markers cyclin E (S-phase), cyclin B (mitosis) and Histone 3 Serine 10 phosphorylation (additional mitotic marker) has to be determined by immunoblot, to deduce the portion of mitotic cells. Depending on the result, two scenarios are conceivable: i) in my S/G2 population, a considerable number of cells already progressed to mitosis, thus allowing origin licensing, and ii) my S/G2 population reflects indeed a time-point after DNA replication, indicating that residual origin licensing could occur earlier than expected.

In *Drosophila melanogaster,* minimal Mcm2-7 loading can occur in absence of cyclin E/Cdk2 activity, while the full complement of Mcm2-7 requires cyclin E/Cdk2 kinase activity during G1/S transition (Powell *et al.* 2015).

The authors propose loading of the full Mcm2-7 complement being dependent on a second wave of Cdc6 expression during G1/S transition (Clijsters and Wolthuis 2014), which is stabilized by cyclin E/Cdk2 kinase activity. Alternatively, minimal loading of Mcm2-7 helicases themselves might promote full Mcm2-7 loading, by a direct Mcm2-7-cyclin E-Cdt1 interaction (Geng *et al.* 2007; Powell *et al.* 2015). These experiments were performed in *Drosophila* Kc and S2 cell lines, treated with several RNAi and chemicals to induce cell cycle arrests at precise stages (early G1, G1 phase, and at G1/S transition). These treatments resulted in an unambiguous phenotype, with minimal Mcm2-7 loading in mitosis/early G1 and full Mcm2-7 complement in G1/S. Subsequent ORC and Mcm2-7 ChIP-seq also clearly revealed ORC and Mcm2-7 co-localization in late mitosis/early G1, whereas little overlap was detected in G1/S. Cell cycle fractionation has the advantage of performing similar experiments without any chemical manipulation that might induce various biases. However, the cell cycle fractions I worked with were not as well defined as specifically arrested cells. Still, I observe similar phenomena: although I did not detect less loaded Mcm2-7 in S/G2 compared to G1, Mcm2-7 positions were more precisely associated to transcriptional regulation and approximated ORC. The fact that the amount of detected Mcm2-7 was not reduced in my experiments in S/G2 can be explained by diluted cell populations, with contaminating G1 and mitotic cell stages. However, closer association of Mcm2-7 to transcriptional regulation and ORC argues for an early stage of Mcm2-7 loading, before spreading occurs.

Taken together, pre-RC ChIP-seq results in S/G2 populations suggest that full pre-RC formation is necessary for Mcm2-7 helicase loading, preferentially occurring at sites of transcriptional regulation. The exact cell cycle stage needs to be further characterized by assessing cyclin expression, but a combination of cyclins already allowing origin licensing is expected. Mcm2-7 positions differ between G1 and S/G2, arguing for Mcm2-7 spreading with onset of G1 phase.

Conclusion

Pre-RC ChIP-seq experiments in G1 and S/G2 cell cycle stages and their analyses at replication initiation/termination zones and regions of active transcription regulation allowed to obtain a genome-wide representation of origin licensing dynamics. Interestingly, Mcm2-7 helicases emerge to be the sole determinant of replication origins (as proposed first in *Xenopus* (Lucas *et al.* 2000; Hyrien, Marheineke, and Goldar 2003; Woodward *et al.* 2006) and recently also claimed in yeast (Das *et al.* 2015) and *Drosophila* (Powell *et al.* 2015)). While pre-RC formation is mandatory for Mcm2-7 double-hexamer loading, further Mcm2-7 spreading prior to S phase argues for replication initiation being independent of complete pre-RC. This scenario would explain low redundancies between ORC positions and active replication data (Miotto, Ji, and Struhl 2016; Petryk *et al.* 2016).

This study describes the first genome-wide ChIP-seq analysis of ORC and Mcm2-7 components in humans. Thereby, these components correlated with zones of active replication initiation, while being specifically depleted from replication termination zones. Actively transcribed genes often constitute termination zones, leading to the hypothesis that active transcription relegates Mcm2-7 helicases to gene-adjacent regions. In S/G2, the specific depletion from termination zones was less pronounced, while association to ORC and active transcription increased. Assuming that S/G2 populations represent a cell stage allowing early replication licensing, these results possibly reflect the initial loading of Mcm2-7 helicases, before spreading occurs. Mcm2-7 spreading might only occur with sufficient Mcm2-7 helicases loaded, and possibly in presence of cyclin E/Cdk2 activity. Also, chromatin anchorage after mitosis might account for Mcm2-7 positions. In eukaryotes, replication timing is established prior to origin selection (Rivera-Mulia and Gilbert 2016b). The timing decision point coincides with re-organization of chromatin into TADs/replication domains after mitosis. If Mcm2-7 helicases were loaded at ORC sites prior to timing decision point and spreading took place after chromatin anchorage, Mcm2-7 density would be higher in gene dense TADs, while it would be lower in gene poor TADs. Assuming that Mcm2-7 dense regions stochastically replicate earlier that Mcm2-7 poor regions, this model would link replication timing to regulation of transcriptions and chromatin environment.

Compartmentalization also localizes heterochromatin TADs towards the nuclear lamina. Heterochromatin is generally replicated late in S phase, indicating that despite a gene-poor environment, efficient origin licensing occurs. In these compartments, chromatin accessibility is most likely not the main determinant of ORC binding and origin licensing, implicating other mechanisms.

One possible candidate is H4K20 methylation, as two methylation stages, H4K20me1 and H4K20me3, have already been shown be relevant for replication regulation. However, the exact mechanism of H4K20 methylation action is elusive.

4.5 ORC ASSOCIATES WITH H4K20ME3 IN HETEROCHROMATIN

Cell cycle-dependent regulation of histone 4 lysine 20 monomethyltransferase PR-Set7 renders H4K20 methylation a promising histone modification to affect DNA replication. Furthermore, stabilization of PR-Set7 (by expressing non-degradable PR-Set7^{PIPmutant} in cells) leads to DNA re-replication, suggesting that origins are re-licensed and re-activated within one cell cycle, if correct regulation of PR-Set7 activity is hampered (Tardat *et al.* 2010; Beck *et al.* 2012). Consequently, H4K20 methylation seems to directly link chromatin regulation and replication licensing and/or activation. For this reason, I performed H4K20me1 and H4K20me3 ChIPs in parallel to pre-RC component ChIPs, to anticipate the role of H4K20 methylation in origin licensing.

4.5.1 H4K20ME1 IS PRESENT IN ACTIVE CHROMATIN WHILE H4K20ME3 LOCALIZES TO HETEROCHROMATIN REGIONS

In collaboration with Eric Julien (Chromatin and Cancer, IRCM, Montpellier) and Jean-Charles Cadoret (Pathology of DNA Replication, Institut Jaques Monod, Paris), H4K20me1 and –me3 were recently assessed genome-wide in U2OS cells. This study revealed H4K20me1 being a histone modification present in active chromatin regions, while H4K20me3 was rather associated with heterochromatin domains (Figure 4.29). These analyses were performed by calculating H3K36me3 (active chromatin) or H3K9me3 (heterochromatin) ChIP-seq read coverage at H4K20me1 or –me3 peaks. Thereby, H4K20me1 sites were increasingly covered by H3K36me3 while H3K9me3 was enriched at H4K20me3 sites (analysis performed by Jean-Charles Cadoret).

H4K20me1 and –me3 ChIPs were performed in parallel to pre-RC component ChIPs in both G1 and S/G2 Raji cells in triplicates (qPCR validation Figure 4.4). Subsequent peak-calling was performed using the HOMER algorithm and the specific "histone" setting. This setting takes the possibility into account that histone modification ChIPs can result in broader peaks. Only peak positions present in all three triplicates were retained. The resulting peaks were around 2000 bp wide for both histone modifications, with both H4K20me1 and –me3 ranging from 200 bp to 88000 bp (Appendix Figure 23 A).

H4K20me1 and –me3 positions in either euchromatic or heterochromatic regions already suggest mutual exclusivity of both histone modifications. When calculating the overlap of H4K20me1 and –me3 peaks, this assumption was confirmed by only ~ 1% of overlapping peaks (Appendix Figure 23 B). This result clearly situated both H4K20me1 and –me3 individually in their respective chromatin environment, without any redundancies.

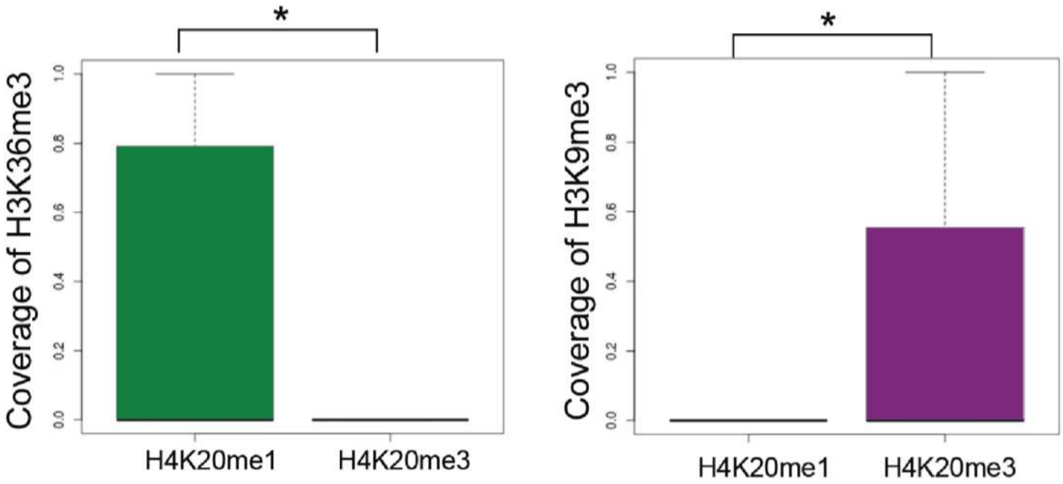


FIGURE 4.29: H4K20ME1 IS PRESENT IN ACTIVE CHROMATIN, WHILE H4K20ME3 CORRESPONDS TO HETEROCHROMATIN. Coverage of either H3K36me3 (active chromatin) or H3K9me3 (silent chromatin) ChIP-seq reads was calculated at H4K20me1/ -me3 peak sites and plotted as boxplot. * indicates statistical significance with p-value < 0.05 (Brustel *et al.*, in preparation).

4.5.2 ORC PREFERENTIALLY ASSOCIATES WITH H4K20ME3

Coverage of pre-RC components was calculated at either H4K20me1 or –me3 sites, to get a first impression of a direct association of either histone modification with pre-RC. All pre-RC components Orc2, Orc3, Mcm3, and Mcm7 were not particularly enriched at H4K20me1 sites (Figure 4.30). Accordingly, also HOMER- and T-PIC-defined ORC, Mcm2-7 and pre-RCs hardly overlapped with H4K20me1 (Appendix Figure 24).

H4K20me3 sites however, were significantly enriched in pre-RC coverage, especially by Orc2 and Orc3, and to a substantial lesser extend by Mcm3 and Mcm7 (Figure 4.31). Also when analyzing the overlaps of ORC, Mcm2-7 and pre-RC with H4K20me3, especially ORC partially overlapped with H4K20me3, while Mcm2-7 and pre-RC showed little association (Figure 4.32). An IGB profile of such an ORC co-localization with H4K20me3 is shown in Appendix Figure 25.

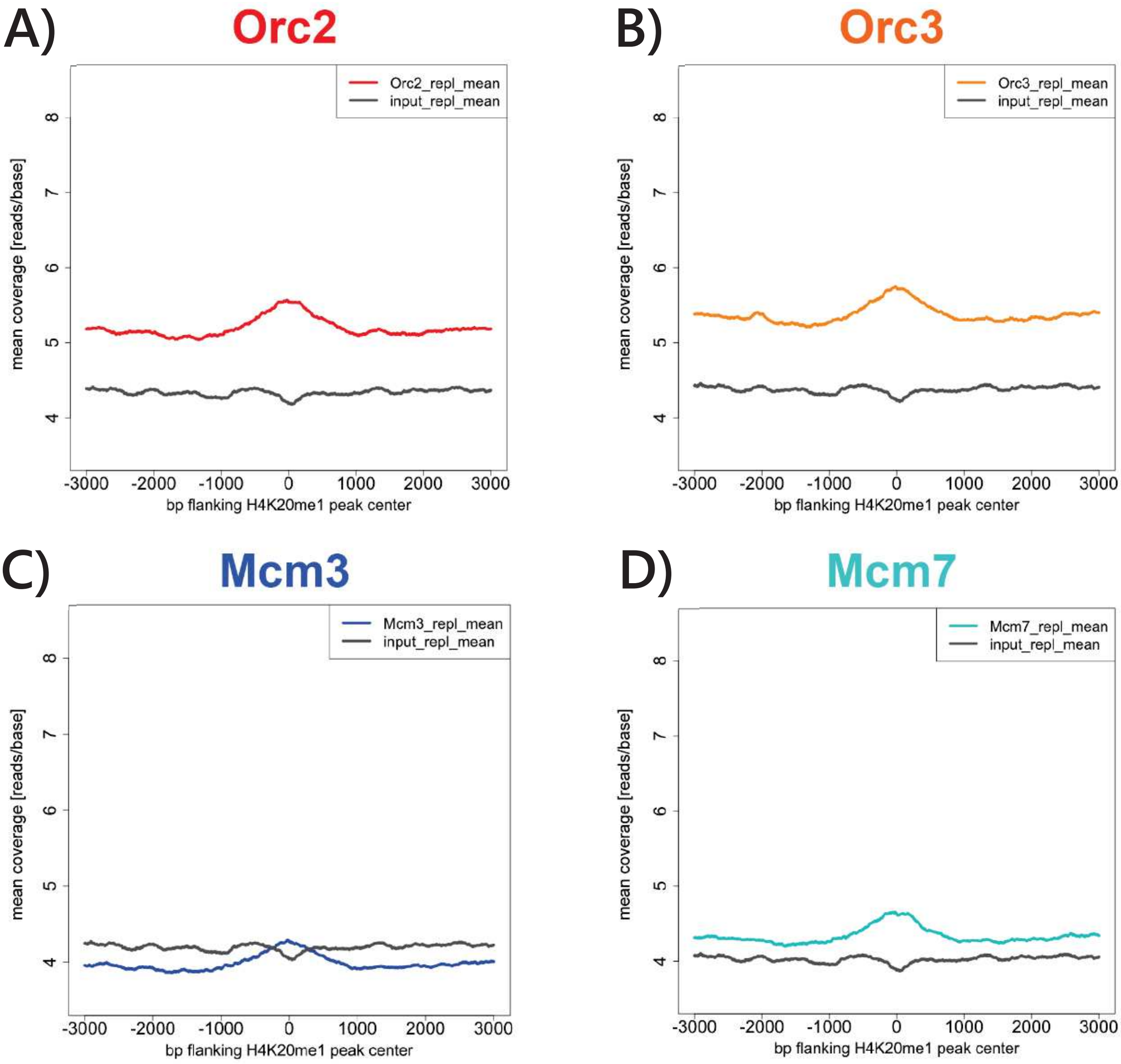


FIGURE 4.30: COVERAGE OF PRE-RC COMPONENTS (G1) AT H4K20ME1 PEAKS. A) MEAN ORC2 COVERAGE. B) MEAN ORC3 COVERAGE. C) MEAN MCM3 COVERAGE. D) MEAN MCM7 COVERAGE. Coverage was calculated as number of reads/base within a 6 kb window around H4K20me1 peak center. Input was plotted as control (grey).

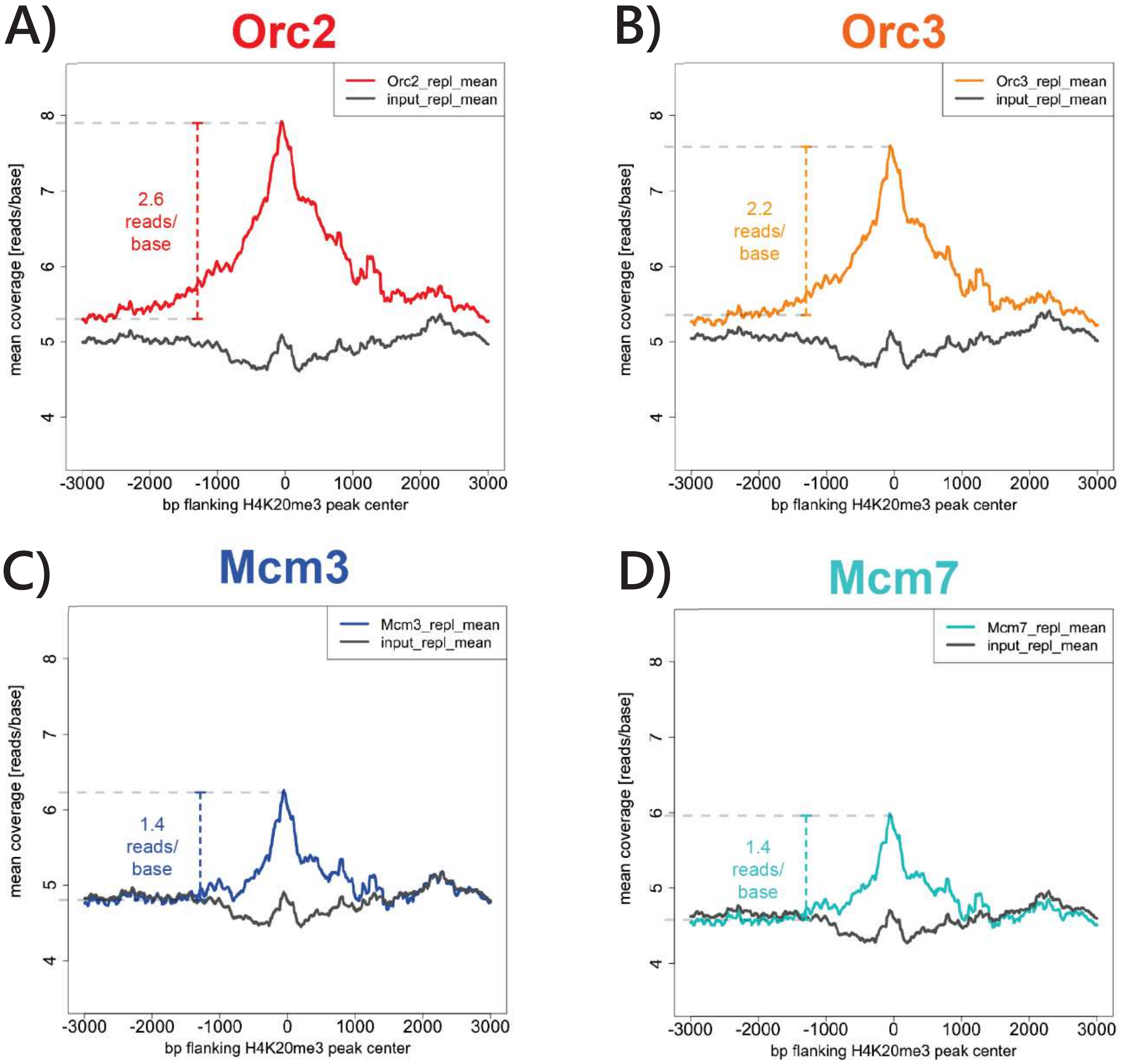


FIGURE 4.31: ORC SHOWED INCREASED COVERAGE AT H4K20ME3 PEAKS. A) MEAN ORC2 COVERAGE. B) MEAN ORC3 COVERAGE. C) MEAN MCM3 COVERAGE. D) MEAN MCM7 COVERAGE. Coverage was calculated as number of reads/base within a 6 kb window around H4K20me3 peak center. Input was plotted as control (grey).

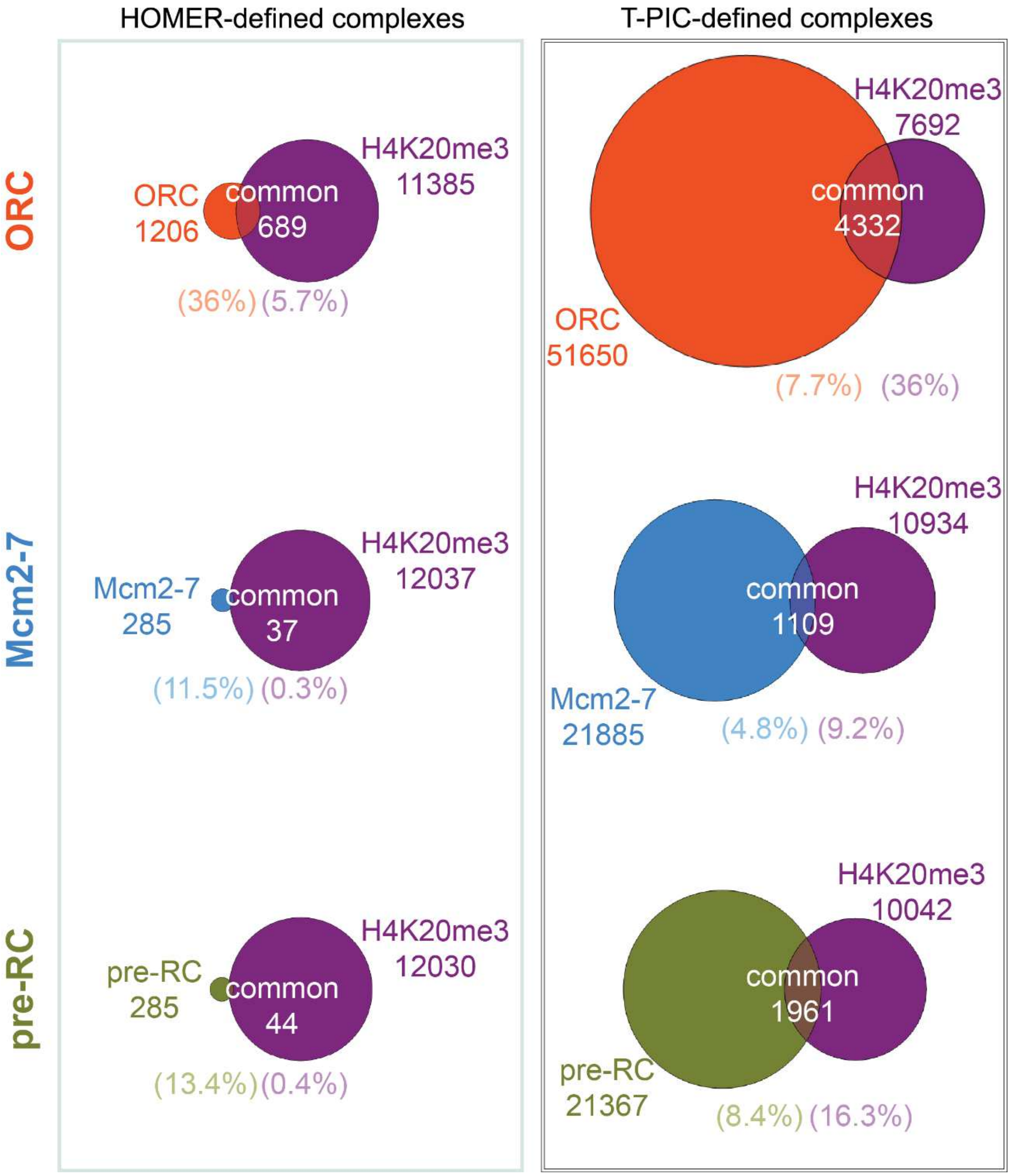


FIGURE 4.32: MOSTLY HOMER- (LEFT) AND T-PIC- (RIGHT) DEFINED ORC OVERLAPPED WITH H4K20ME3. Left: HOMER-defined complexes, right: T-PIC-defined complexes. Venn diagram of overlap between HOMER- and T-PIC-defined ORC, Mcm2-7 and pre-RCs with H4K20me3 peaks. Overall counts are indicated. The percentage of overlapping proportions are specified below in brackets.

4.5.3 NO DISTRIBUTIONAL CHANGE WAS OBSERVED IN S/G2

H4K20 methylation ChIPs analyzed so far were performed in G1 cell cycle stage. As PR-Set7 is cell cycle dependently regulated, with degradation during S phase and a subsequent decrease of H4K20me1, I was wondering how H4K20 methylation peaks and the observed associations were modulated in S/G2.

In general, detected number of H4K20me1 and –me3 peaks decreased in S/G2 compared to G1. This observation was more prominent for H4K20me1 (from 7820 peaks in G1 to 1700 peaks in S/G2), than for H4K20me3 (from 11709 peaks in G1 to 5560 peaks in S/G2). Thereby, most of the peaks detected in S/G2 overlapped with G1 peaks (> 84%, Appendix Figure 26), arguing for a dilution of H4K20 methylation during replication, which led to weaker ChIP signals, presumably below the threshold for peak-calling. This was also reflected by a decrease of H4K20 methylation coverage at their own respective G1 peaks (Appendix Figure 27). Again, H4K20me1 seemed to be more affected than H4K20me3, where no obvious coverage change was detected.

Neither coverage of any pre-RC component, nor peak positions relative to either H4K20 methylation changed considerably in S/G2 (data not shown), indicating that no major licensing mechanism relies on the cell-cycle distribution of H4K20me1/ -me3.

POSSIBLE MECHANISMS UNDERLYING ORIGIN LICENSING IN HETEROCHROMATIN

Previous reports suggest that PR-Set7 and subsequent induction of H4K20me1 undoubtedly affects origin licensing and activation. An extensive genome-wide SNS-seq study correlating active replication with histone modifications claimed H4K20me1 associating with 50% of initiation sites, being one of the potential key regulators of replication (Picard *et al.* 2014). However, pre-RC coverage at H4K20me1 sites was not majorly enriched. This is in line with other genome-wide studies that did not find any associations between H4K20me1 and replication (Fu *et al.* 2013; Cayrou *et al.* 2015; Smith *et al.* 2016; Miotto, Ji, and Struhl 2016). By contrast, ORC coverage was clearly enriched at H4K20me3 sites, Mcm3 and Mcm7 coverage to a lesser extent. This is in accordance with ORC directly interacting with H4K20me2/3, most probably through the BAH domain of Orc1 (Kuo *et al.* 2012; Beck *et al.* 2012). ORC interaction with heterochromatin has already been reported repeatedly in yeast and *Drosophila* (Micklem *et al.* 1993; Pak *et al.* 1997; Shareef, Badugu, and Kellum 2003; Leatherwood and Vas 2003; Shen *et al.* 2010). Thereby, ORC was attributed a chromatin silencing function independent of origin licensing (Dillin and Rine 1997; Leatherwood and Vas 2003). Regarding the low Mcm2-7 coverage enrichment, this might be an explanation for ORC – but not Mcm2-7 – associating with H4K20me3. However, dynamics of Mcm2-7 loading might also be different in heterochromatin and little Mcm2-7 is still sufficient for efficient replication initiation. Consequently, I examined whether ORC association to heterochromatin H4K20me3 was indeed linked to origin replication licensing and activation.

4.6 H4K20me3 is Necessary for Origin Licensing and Activation in Heterochromatin

The association of ORC with H4K20me3 is not necessarily linked to replication licensing function of ORC. Mcm2-7 helicases seemed to be less present at H4K20me3 sites, than ORC (Figure 4.31), eventually arguing for a heterochromatin organization function of ORC. I used the replication of the well-characterized autosomal plasmid system of the Epstein-Barr virus latent origin *oriP* (herein after called FR-DS, Hammerschmidt and Sugden 2013) to functionally validate the role of ORC and H4K20me3 association. Contrarily to genome-wide analysis or genomic integration sites, this model system with defined genetic background allows the analysis of single aspects in the replication initiation process. The strategy was to specifically induce H4K20me3 at a distinct targeting site on the plasmid and to assess the effect on licensing and replication activity by ChIP-qPCR and plasmid abundance. For autonomous replication and segregation of FR-DS plasmids, the EBV viral protein EBNA1 needs to be expressed by the cells (Figure 4.33 A). By binding of EBNA1 to the FR element, plasmids are tethered to the host chromatin conferring mitotic stability. EBNA1 binding to the DS element recruits ORC, which leads to the formation of an efficient internal origin of replication. Different reporter plasmids were generated: Introducing an upstream activation sequence (UAS) of the Gal4-UAS targeting system downstream of FR allows the specific targeting of Gal4(-fusion) proteins. DS can be removed or replaced by a 300 bp fragment of the human endogenous origin ori^{RDH}, proven to also support autonomous plasmid replication, however less efficiently than DS (Gerhardt *et al.* 2006, described as ori6).

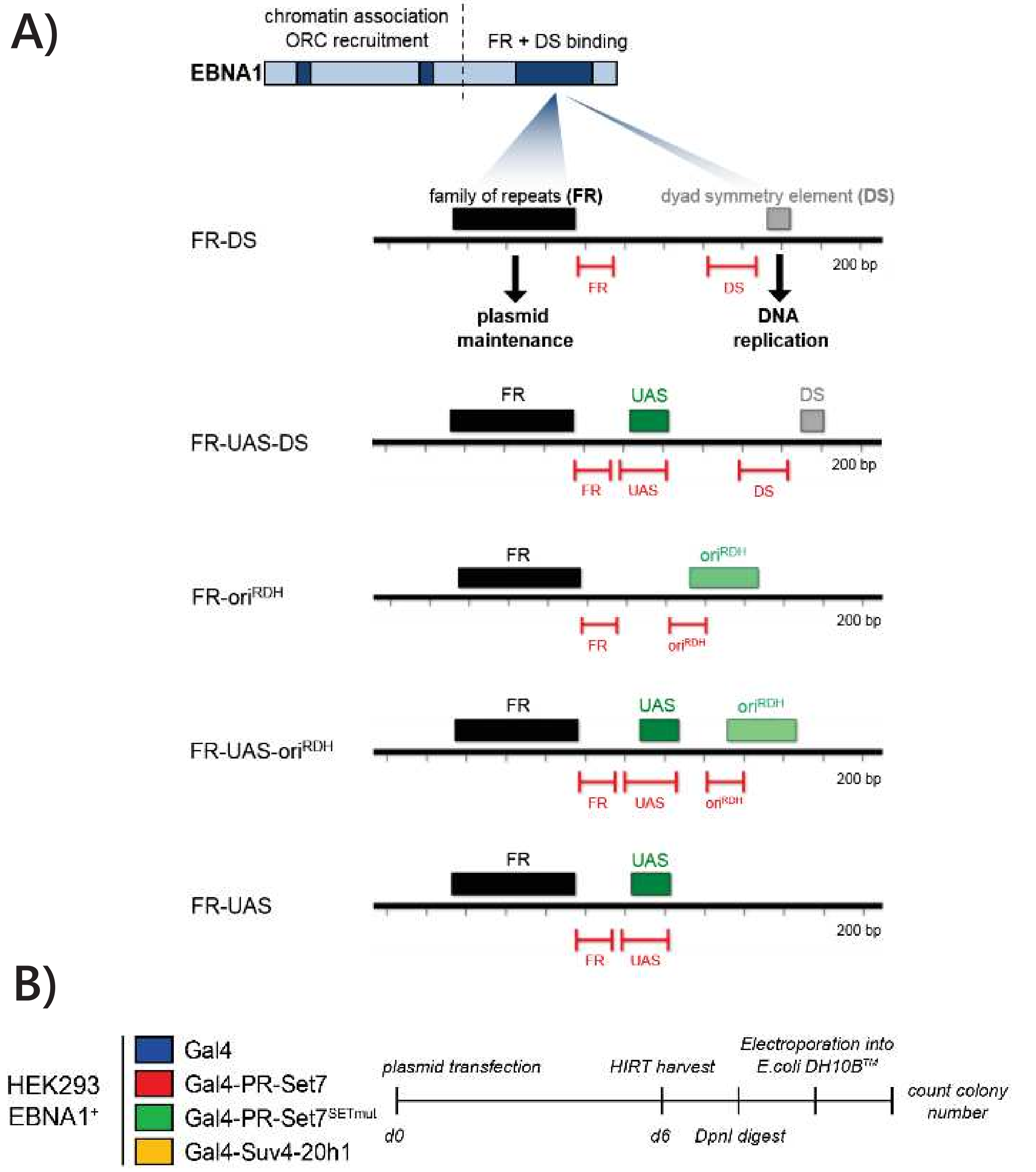


FIGURE 4.33: SCHEMATIC REPRESENTATION OF EXPERIMENTAL SETUP. A) EBV-DERIVED AUTOSOMAL REPORTER PLASMID SYSTEM. Five EBV-derived reporter plasmids are segregated throughout cell division by EBNA1-mediated tethering to host chromatin. Introduction of UAS sequence allows specific targeting of Gal4(-fusion) proteins. The internal replicator DS was removed or replaced by the human endogenous origin ori^{RDH}. DNA fragments amplified by qPCR are indicated in red. B) PLASMID ABUNDANCE EXPERIMENTS. Experiments were conducted in HEK293 cells stably expressing EBNA1 and the indicated Gal4(-fusion) proteins. 1µg of reporter plasmids were transfected in the cells, kept replicating for 6 days and isolated from the cells by the HIRT protocol, specifically enriching for low molecular weight DNA. DpnI digest removes bacterial-derived input plasmids and DpnI-resistant plasmids were electroporated in *E.coli* DH10B. Bacterial colony quantification is a direct measure of plasmid numbers.

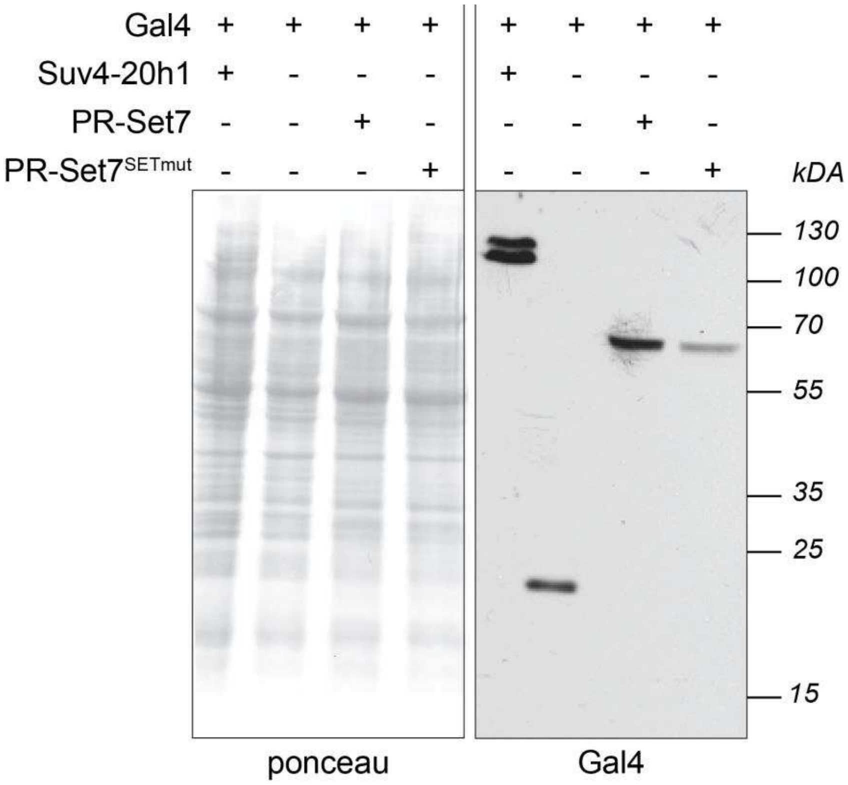


FIGURE 4.34: GENERATION OF HEK293 EBNA1$^+$ GAL4 (-FUSION) CELL LINES. pcDNA3_Zeo expression plasmids carrying the Gal4(-fusion) protein cassettes (obtained from E. Julien, IRMB, Montpellier) were linearized and transfected into HEK293 EBNA1$^+$ cells. After selection and clonal expansion, depicted clones were chosen for further experiments. Ponceau stain is shown as loading control, immunoblot was performed using an anti-Gal4 antibody.

4.6.1 SUV4-20H1 TARGETING UNSPECIFICALLY INDUCES H4K20ME3

The first and very easy reasoning to define H4K20me3 action in replication processes was to simply target the H4K20me2/3 generating histone methyltransferase Suv4-20h1 to the FR-UAS-ori^{RDH} reporter plasmid. Consequently, either Gal4 or Gal4-Suv4-20h1 expression cassettes were integrated into HEK293 EBNA1$^+$ cells and their expression was tested by immunoblot (Figure 4.34, first two lanes). Interestingly, integration of Gal4-Suv4-20h1 resulted in a doublet signal around 120 kDa, suggesting a post-translational modification of this fusion protein (expected size ~ 100 kDa).

Induction of Suv4-20h1-mediated H4K20me3 was verified by H4K20me3 ChIP experiments for the FR-ori^{RDH} and FR-UAS-ori^{RDH} plasmids, followed by qPCR. While Gal4 and Gal4-Suv4-20h1 were specifically targeted to UAS (Gal4, Figure 4.35, left panel), Gal4-Suv4-20h1 induced H4K20me3 spread over the plasmid, even when no targeting site is present (Figure 4.35, right panel, FR-ori^{RDH} vs. FR-UAS-ori^{RDH}). Consequently, plasmid abundance experiments did not result in any effect of Gal4-Suv4-20h1 targeting (Appendix Figure 28), possibly due to heterochromatinization of both reporter plasmids.

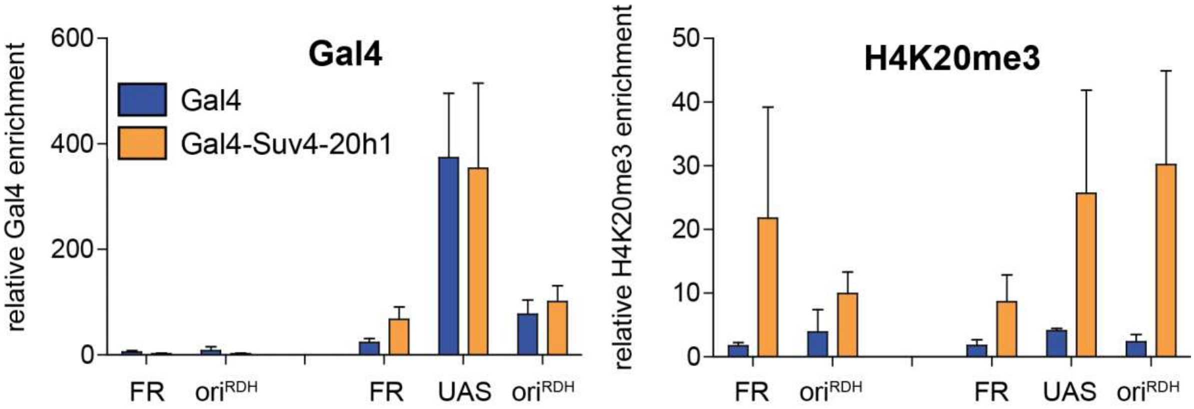


FIGURE 4.35: SUV4-20H1 TARGETING TO FR-ORI^{RDH} AND FR-UAS ORI^{RDH} UNSPECIFICALLY INDUCES H4K20ME3 ALL OVER THE PLASMIDS. ChIP-qPCR analyses at FR, UAS and ori^{RDH} sequences of FR-ori^{RDH} or FR-UAS-ori^{RDH} plasmids transfected in the indicated cell lines. Fold enrichments relative to IgG. Data are means ± SEM (n=3).

Reason for the observed unspecific induction of H4K20me3 can be the heavy overexpression of Gal4-Suv4-20h1. Assuming that also the modified, higher migrating form of Suv4-20h1 is functional, expression is doubled compared to Gal4 or Gal4-PR-Set7 proteins (Figure 4.34). Although the cell line itself did not exhibit any abnormalities concerning cell growth or morphology, it seems that Gal4-Suv4-20h1 is inducing H4K20me3 completely independent from targeting to UAS. This renders the approach of Gal4-Suv4-20h1 targeting not suitable for our purposes.

I refrained from producing a lower expressing cell line and decided to instead target H4K20 monomethyltransferase PR-Set7 to the plasmids. Induction of H4K20me1 will be converted in H4K20me2/3 by endogenous Suv4-20h1/2.

4.6.2 PR-SET7 TARGETING LEADS TO INDUCTION OF H4K20ME1, CONVERSION INTO H4K20ME3 AND PRE-RC FORMATION

Either FR-ori^{RDH}/ FR-UAS-ori^{RDH} or FR-DS/ FR-UAS-DS reporter plasmids were transfected in cell lines expressing Gal4, Gal4-PR-Set7 or a methylation-deficient N469A/H470A SET mutant of PR-Set7, Gal4-PR-Set7^{SETmut} (for integrated protein expression, see Figure 4.34, lanes 2-4).

Efficient induction of H4K20me1 by PR-Set7 and conversion to H4K20me3 was assessed by ChIP-qPCR. Furthermore, pre-RC formation was followed by the same technique, analyzing the specific pre-RC sub-component Mcm3. Indeed, Gal4-PR-Set7 targeting to FR-UAS-ori^{RDH} plasmid induced H4K20me1, which was subsequently converted in H4K20me3 (Figure 4.36).

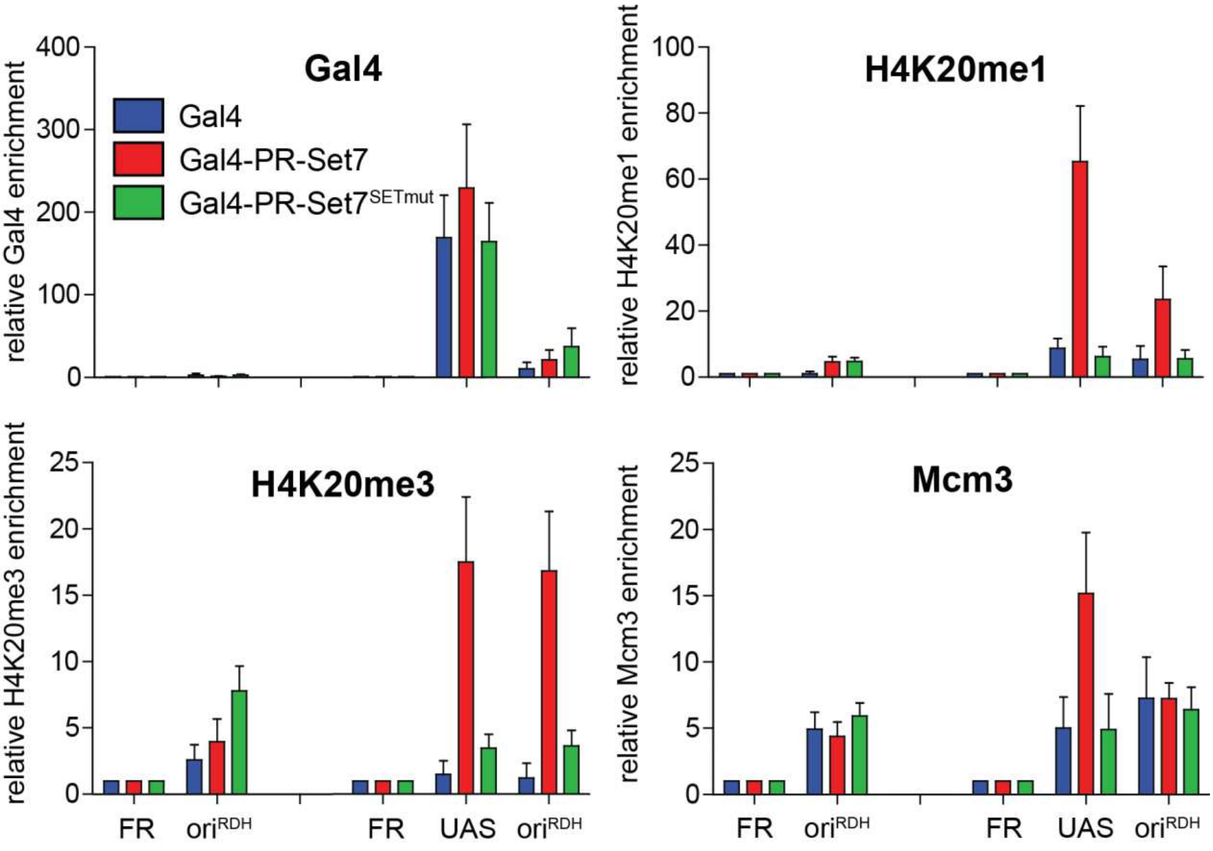


FIGURE 4.36: PR-SET7 TARGETING TO FR-UAS-ORI^{RDH} LED TO H4K20ME1 INDUCTION, CONVERSION INTO H4K20ME3 AND PRE-RC FORMATION. ChIP-qPCR analyses at FR, UAS and ori^{RDH} sequences of FR-ori^{RDH} or FR-UAS-ori^{RDH} plasmids transfected in the indicated cell lines. Fold enrichments relative to IgG and FR locus. Data are means ± SEM (n=4).

This was not observed when Gal4 or the methylation-deficient PR-Set7^{SETmut} were targeted to the reporter plasmids. Interestingly, both H4K20me1 and –me3 were also detected at ori^{RDH} when PR-Set7 was targeted, presumably due to the close proximity of UAS and ori^{RDH} primer positions (154 bp ~ 1 nucleosome). More importantly, endogenous Mcm3 as member of the pre-RC was detected at UAS in the Gal4-PR-Set7 cell line only, demonstrating pre-RC formation after H4K20me1/3 induction (Figure 4.36). Similar results were obtained when the respective Gal4, Gal4-PR-Set7 and Gal4-PR-Set7^{SETmut} cell lines were transfected with *oriP* reporter plasmids FR-DS and FR-UAS-DS (Figure 4.37).

It has already been shown that PR-Set7 targeting to a UAS integrated in the human genome leads to pre-RC formation (Tardat *et al.* 2010). I validated the autonomous plasmid system by reproducing these results. However, recruitment of pre-RC components is not necessarily sufficient for initiation of DNA replication (Blow, Ge, and Jackson 2011). Consequently, I assessed origin activity of the plasmids by measuring plasmid abundance.

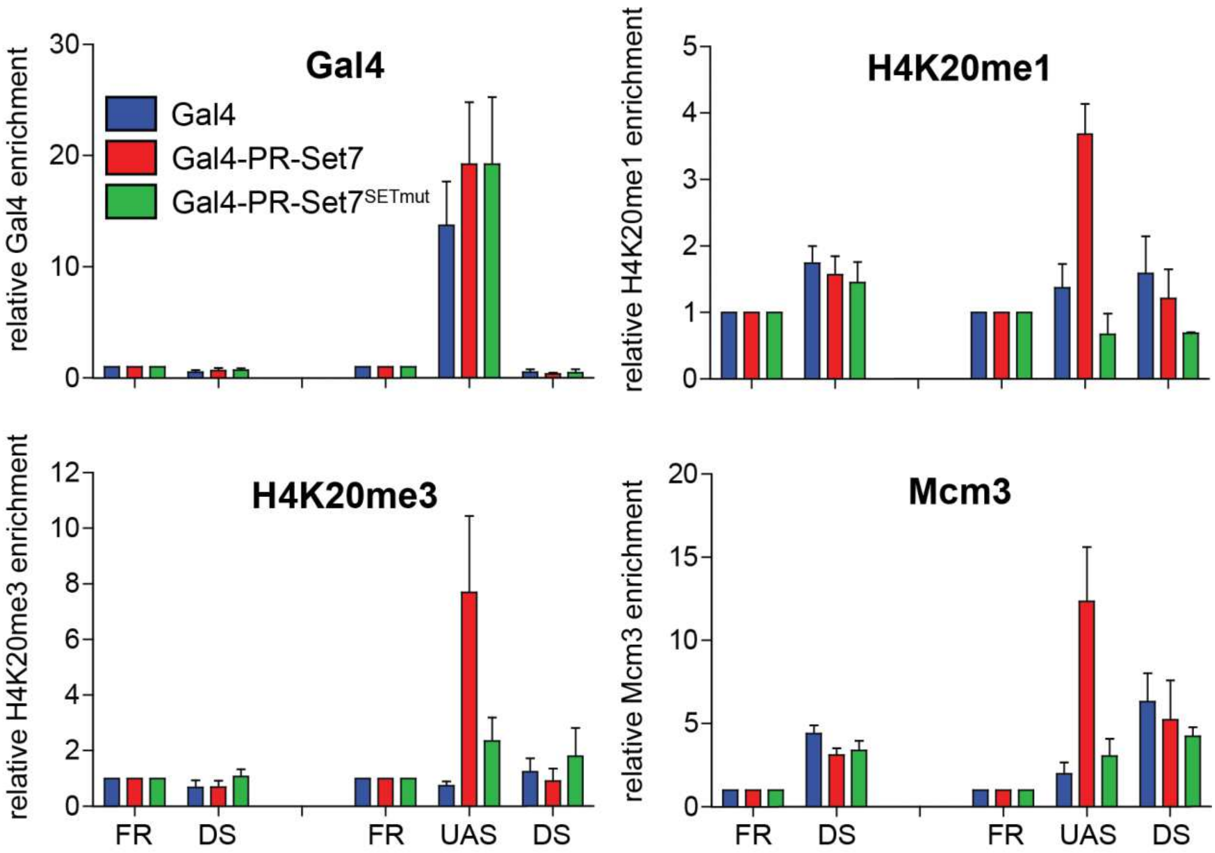


FIGURE 4.37: PR-SET7 TARGETING TO FR-UAS-DS LED TO H4K20ME1 INDUCTION, CONVERSION INTO H4K20ME3 AND PRE-RC FORMATION. ChIP-qPCR analyses at FR, UAS and DS sequences of FR-DS or FR-UAS-DS reporter plasmids transfected in the indicated cell lines. Fold enrichments relative to IgG and FR locus. Data are means ± SEM (n=4).

4.6.3 PR-SET7 TARGETING ENHANCES PLASMID REPLICATION EFFICIENCIES

Principal of the plasmid abundance assay is the measurement of reporter plasmid replication efficiencies within a fixed time frame (procedure described in Figure 4.33 B). Shortly, I transfected reporter plasmids (either FR-ori^{RDH}/ FR-UAS-ori^{RDH} or FR-DS/ FR-UAS-DS) in Gal4, Gal4-PR-Set7 or Gal4-PR-Set7^{SETmut} cell lines. Cells with transfection efficiencies greater 70% were grown for 6d before low molecular weight DNA was harvested, *DpnI* digested and transformed into *E.coli*. *DpnI* digest removes bacterial-derived input plasmids. The appearance of bacterial colonies from *DpnI*-resistant plasmids is a direct measure of plasmid replication in mammalian cells. Because reporter plasmid replication efficiencies varied between cell lines, replication efficiencies of the control reporter plasmid without UAS targeting site were arbitrarily defined as 1. Replication efficiencies of the UAS-containing plasmids were calculated compared to control in each cell line and the result is depicted in Figure 4.38.

For the FR-ori^{RDH}/ FR-UAS-ori^{RDH} plasmid pair, FR-UAS-ori^{RDH} replicated about 3.5-times more than control when Gal4-PR-Set7 was targeted to the plasmid (Figure 4.38 A). This effect was solely attributable to PR-Set7 methylation activity, as no effects were observed when either Gal4 or Gal4-PR-Set7^{SETmut} were targeted. Plasmid abundance experiments with the

FR-DS/ FR-UAS-DS plasmid pair resulted in 2-fold enhanced replication efficiencies when Gal4-PR-Set7 was targeted (Figure 4.38 B). Replication increase of only 2-fold (compared to 3.5-fold in ori^{RDH} context) might be due to the high intrinsic replication competence of DS.

These results suggest that induction of H4K20me1 by PR-Set7 targeting, subsequent conversion into H4K20me3 and pre-RC formation led to replication initiation, in addition to intrinsic plasmid replication competences. Indeed, origin competence likely involves combination of different features ranging from sequence composition to higher-order chromatin structure (Méchali 2010). Thus, these results do not allow to distinguish between H4K20me1/3 increased origin activation as an enhancement of intrinsic origin activity or of additional origin formation. To pursue this question, the same experiments were performed with a reporter plasmid devoid of any replication competence.

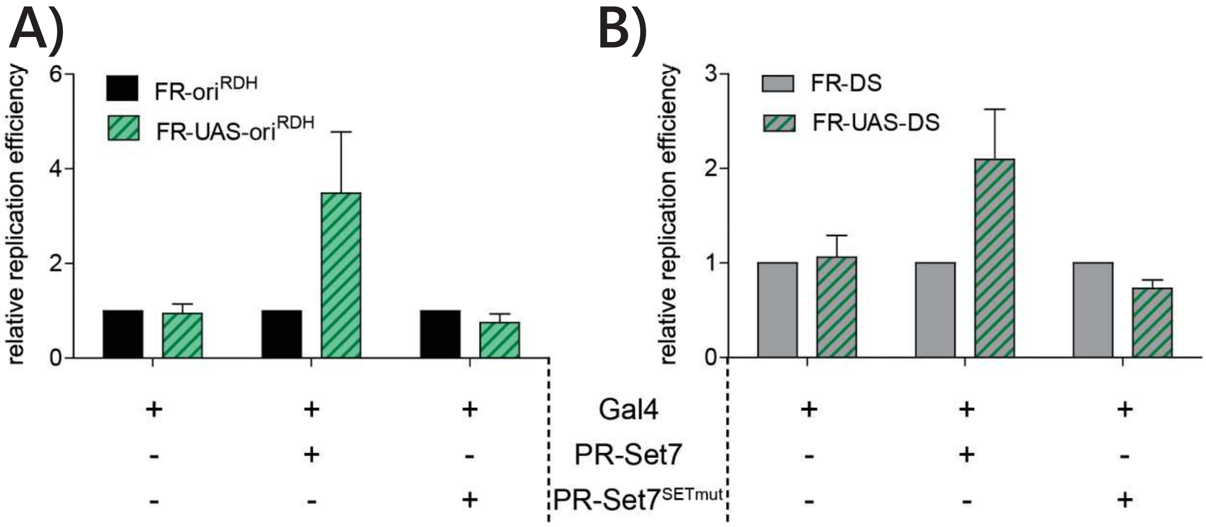


FIGURE 4.38: PR-SET7 TARGETING ENHANCED INTRINSIC PLASMID REPLICATION ACTIVITY. A) QUANTIFICATION OF FR-ORI^{RDH} AND FR-UAS-ORI^{RDH} PLASMIDS. B) QUANTIFICATION OF FR-DS AND FR-UAS-DS PLASMIDS. Plasmid abundance assays in HEK293 EBNA1^{+} Gal4/ Gal4-PR-Set7/ Gal4-PR-Set7^{SETmut} cell lines as indicated: Control plasmid (without UAS) replication efficiencies were arbitrarily defined as 1. Data are means ± SEM (n=4).

4.6.4 PR-SET7 TARGETING INDUCES PLASMID REPLICATION

To test whether PR-Set7 targeting can also lead to the induction of DNA replication, I removed the DS element from FR-UAS-DS plasmid and assessed the resulting replication efficiency compared to the strong replicating FR-DS reporter (Figure 4.39). While FR-UAS reporters were hardly detectable in Gal4 and Gal4-PR-Set7^{SETmut} cell lines, FR-UAS plasmids clearly replicated in Gal4-PR-Set7 cell lines. The induced replication efficiency was about 1/3 as strong as the intrinsic DS activity and 4-fold increased to FR-UAS plasmid replication in Gal4 or Gal4-PR-Set7^{SETmut} cell lines. A direct comparison of all reporter plasmid replication efficiencies is depicted in Appendix Figure 29.

ChIP experiments equally revealed H4K20me1/me3 being present at UAS, but due to the low plasmid abundance, it was challenging to conclusively perform Mcm3 ChIPs (Appendix Figure 30). Still, pre-RC formation is a prerequisite for replication activity.

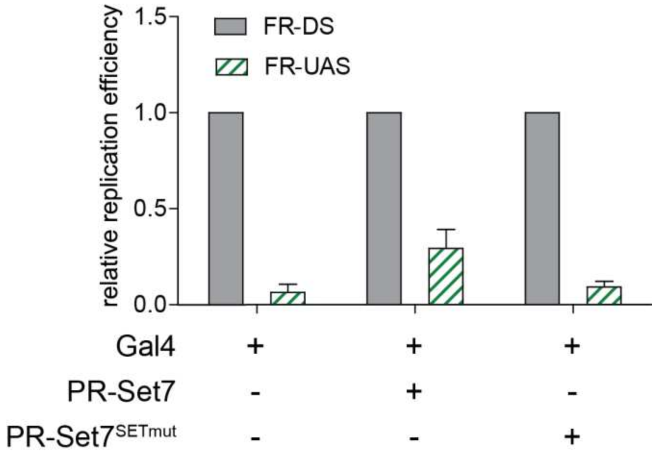


FIGURE 4.39: PR-SET7 TARGETING INDUCED PLASMID REPLICATION. QUANTIFICATION OF FR-DS AND FR-UAS PLASMIDS. Plasmid abundance assays in HEK293 EBNA1^{+} Gal4/ Gal4-PR-Set7/ Gal4-PR-Set7^{SETmut} cell lines as indicated: FR-DS replication efficiency was arbitrarily defined as 1. Data are means ± SEM (n=3).

In conclusion, induction of H4K20me1/3 leads to origin licensing and activation. However, no conclusion can be drawn whether conversion into H4K20me3 is needed for these processes.

4.6.5 REPLICATION LICENSING AND ACTIVATION DEPEND ON H4K20ME3

There is a lot of evidence that conversion from H4K20me1 to H4K20me3 is indeed needed for origin licensing and activation (Beck *et al.* 2012). Also my pre-RC ChIP-seq experiments imply the necessity for H4K20me3, as H4K20me1 is only marginally covered by pre-RC sub-components (Figure 4.30), while especially ORC seems to have a preference for H4K20me3 (Figure 4.31).

Eric Julien (Chromatin and Cancer, IRCM, Montpellier), a close collaboration partner, also investigated the interplay of H4K20me3 and replication origins and showed that H4K20me3 is necessary for proper origin licensing. They performed replication timing experiments in MEF cells derived from Suv4-20h2 $^{-/-}$, Suv4-20h1 $^{-/flox}$, Cre-ER embryos. The remaining Suv4-20h1 floxed allele can be deleted by adding 4-Hydroxytamoxifen (4-OHT), leading to absence of H4K20me2/3 and an increase in H4K20me1 (immunoblot 4d after 4-OHT treatment, Figure 4.40, (Schotta *et al.* 2008)). Comparing replication timing in wild type MEF and Suv4-20h knock-out MEF revealed a delay of 29% of late replication timing regions while some early replicating regions were even slightly advanced (red vs. blue arrows, Figure 4.40 A).

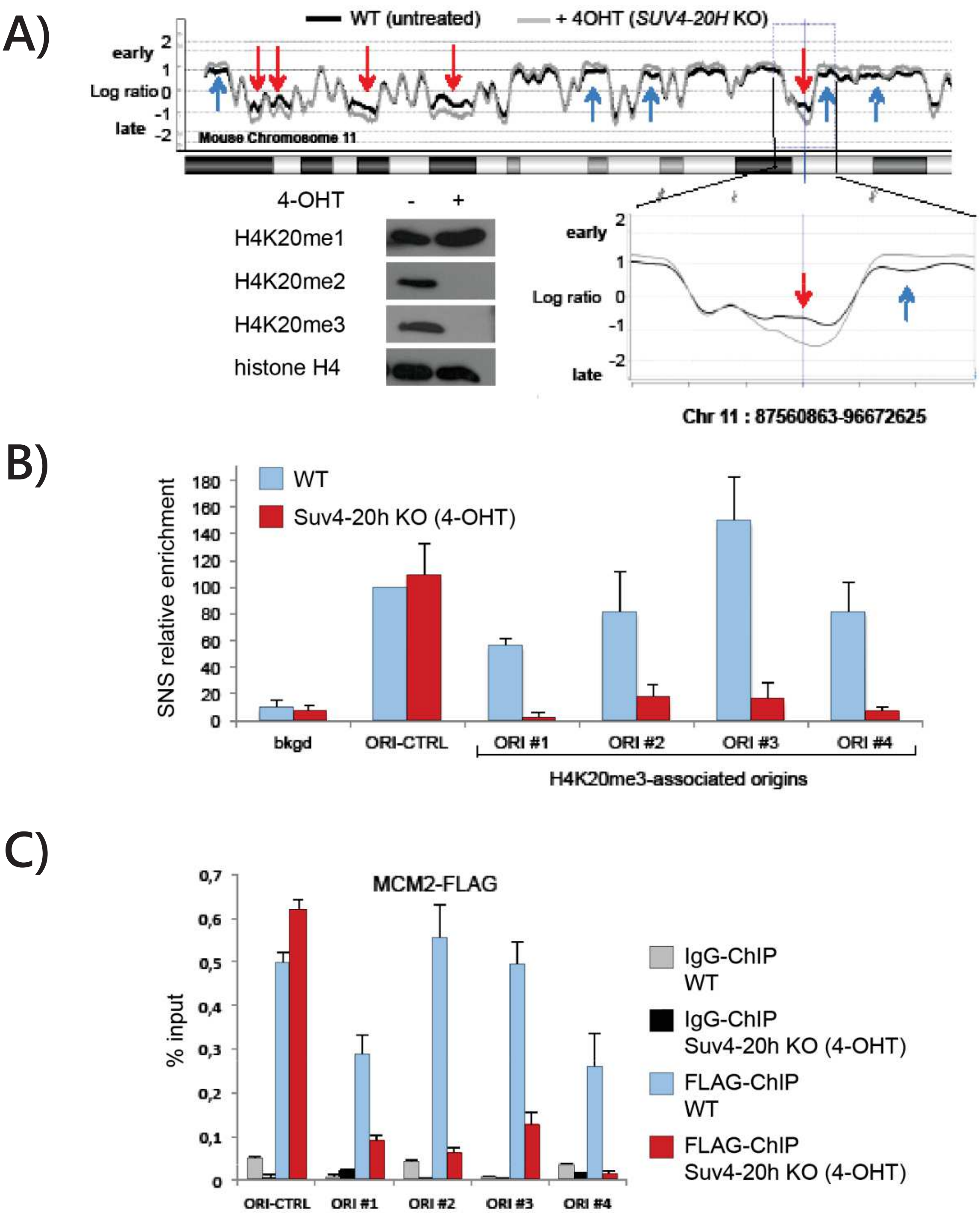


FIGURE 4.40: ABSENCE OF H4K20ME2/3 IMPAIRS ORIGIN LICENSING AND ACTIVATION IN DEFINED DOMAINS IN MEF CELLS. A) REPLICATION TIMING PROFILES ON CHROMOSOME 11 IN SUV4-20H1 FLOX/-, SUV4-20H1 -/- MEFS, TREATED OR NOT WITH 4-OHT. Lower left panel: immunoblot confirmation of H4K20me2/3 absence. Lower right panel: replication timing in region [87560863-96672625]. Red arrows indicate delayed late replicating domains, blue arrows point to advanced early replicating domains. B) RELATIVE SNS ENRICHMENT of four H4K20me3 late-firing origins compared to control early-firing origin. H4K20me3-associated origins from chromosome 11 were quantified in relation to an early-firing origin in untreated and 4-OHT treated MEF cells. Data are means ± SEM (n=3). C) CHIP-QPCR ANALYSIS at the control early-firing origin and at four H4K20me3-associated origins from chromosome 11 in untreated and 4-OHT treated MEFs expressing FLAG-tagged Mcm2. Y-axis represents the ration of immunoprecipitate to input. (Figure adapted from Brustel *et al.* in preparation)

H4K20me3 is mostly present in heterochromatin, which is known to be late replicating. Specific measurement of origin activity in some of these late replicating domains delayed by Suv4-20h1/2 knock-out (H4K20me3-associated origins) revealed severe reduction of origin efficiency upon Suv4-20h1/2 knock-out (Figure 4.40 B). Following ChIP-qPCR analysis at the same H4K20me3-associated origins showed a prominent decrease in pre-RC formation (assessed by Mcm2-FLAG ChIP), when H4K20me2/3 was absent (Figure 4.40 C). Consequently, delay of replication timing upon Suv4-20h1/2 knock-out results from the inability to properly license origins in regions that depend on H4K20me3. These regions might instead be replicated passively. These results already indicate the necessity of H4K20me3 for origin licensing and activity. To confirm these observation using the plasmid system with defined genetic background, I need the possibility to precisely block conversion of H4K20me1 to –me3 by inhibiting Suv4-20h1/2.

The inhibitor compound A-196 specifically inhibits Suv4-20h1/2 methylation activity and results in a complete loss of H4K20me2/3, with a concomitant increase of H4K20me1, while other trimethylated histone states remained unaffected (examined by immunoblot, Figure 4.41). To directly evaluate A-196 effect on H4K20me1/3 induced plasmid replication, I transfected HEK293 EBNA1^{+} Gal4-PR-Set7 cells with FR-ori^{RDH}/ FR-UAS-ori^{RDH} reporter plasmids and directly compared A-196 treated with untreated cells. ChIP-qPCR analysis revealed induction of H4K20me1 and conversion to H4K20me3 in the untreated cell population, as expected, while A-196 treated cells were reduced in H4K20me3 (Figure 4.42). Analyzing the presence of endogenous Mcm3 revealed failure of pre-RC formation at the UAS in A-196 treated cells (Figure 4.42). The incapacity to properly form pre-RCs at the UAS also led to reduction of plasmid replication efficiency to the level of the intrinsic origin ori^{RDH} (Figure 4.43). In conclusion, experiments using the autonomous plasmid system showed that H4K20me3 is definitely needed for proper origin licensing and consequently also for replication activity. Furthermore, genome-wide replication timing experiments revealed specific heterochromatic regions that depend on H4K20me3, while replication in other regions is presumably regulated through different mechanisms.

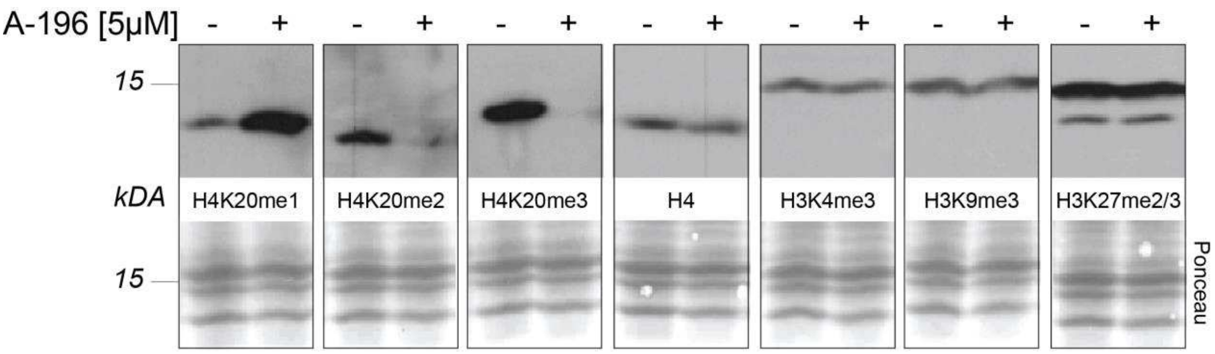


FIGURE 4.41: IMMUNOBLOT CONFIRMATION OF A-196 COMPOUND INHIBITING H4K20ME2/3. HEK293 EBNA1+ Gal4-PR-Set7 cells were treated or not with 5µM A-196 for 6d. Indicated antibodies were used for detection of the respective histone modification. Ponceau stain of histones served as loading control.

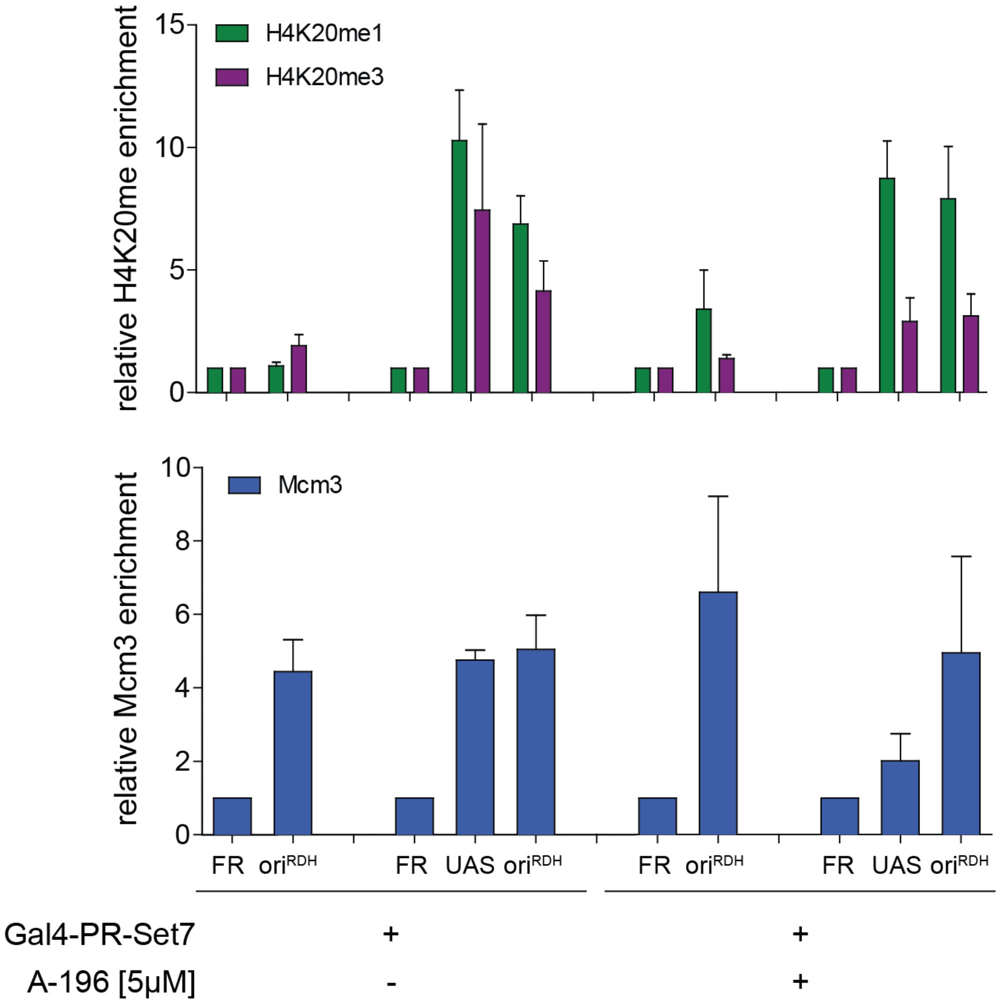


FIGURE 4.42: PR-SET7 TARGETING IN PRESENCE OF A-196 LED TO REDUCED H4K20ME3 LEVELS AND PRE-RC FORMATION COMPARED TO UNTREATED CELLS. ChIP-qPCR analysis at FR, UAS and oriRDH sequences of FR-oriRDH or FR-UAS-oriRDH reporter plasmids transfected in HEK293 EBNA1+ Gal4-PR-Set7 cells treated or not with 5µM A-196. Fold enrichment relative to IgG and FR locus. Data are means ± SEM (n=3).

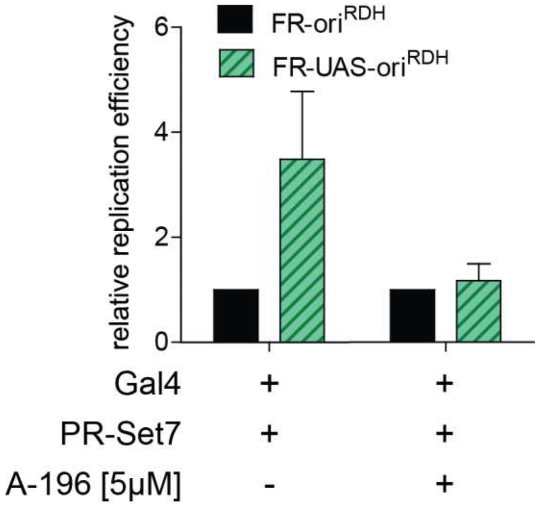


FIGURE 4.43: ENHANCED REPLICATION DURING PR-SET7 DEPENDS ON H4K20ME3. QUANTIFICATION OF FR-ORI^{RDH} AND FR-UAS-ORI^{RDH} PLASMIDS. Plasmid abundance assays in HEK293 EBNA1^{+} Gal4-PR-Set7 cells treated or not with 5µM A-196 as indicated: FR-ori^{RDH} replication efficiency was arbitrarily defined as 1. Data are means ± SEM (n=3).

H4K20ME1 IS LIKELY NOT DIRECTLY LINKED TO REPLICATION ORIGIN LICENSING

Cell cycle regulation of the H4K20 monomethyltransferase PR-Set7 and resulting catalyzation of H4K20me1 during late G2/M led to the hypothesis of this histone modification being involved in DNA replication regulation. My ChIP-seq experiments of H4K20me1 and –me3 revealed both methylation marks being mutually exclusive. H4K20me2 was not assessed, as the abundance of this mark argues against a possible role in DNA replication regulation (Pesavento *et al.* 2008). Together with Eric Julien (Chromatin and Cancer, IRCM, Montpellier), we identified the conversion from H4K20me1 to H4K20me3 to influence origin licensing, origin activation efficiency and replication timing, thus representing the first histone modification identified so far, that impacts on all levels of DNA replication regulation. Thereby, H4K20me3 affects origin licensing, consequently impairing origin formation at the very first step. This holds true for artificial induction of H4K20 methylation states on plasmids with defined genetic background, as well as for specific genomic loci with less defined chromatin context. Nevertheless, in both cases, origin licensing and activation entirely depend on the presence of H4K20me3.

While knock-out of PR-Set7 is lethal in mammalians, absence of both Suv4-20h1 and -h2 displays only minor cell cycle defects (Oda *et al.* 2010; S. Wu and Rice 2011; Beck *et al.* 2012; Schotta *et al.* 2008). Indeed, genome-wide timing experiments only revealed a delay of some late replicating domains when Suv4-20h1/2 were missing. This indicates that while replication origin licensing and activation is perturbed in absence of H4K20me2/3, the concerned regions are still replicated, most likely passively. These results imply that the severe phenotype of PR-Set7 knock-out presumably depend on other functions than replication regulation. It has recently been reported, that unmethylated H4K20 (H4K20me0) marks post-replicative chromatin in G2 and is read by the H3-H4 histone chaperone TONSL-MMS22L, which is implicated in DNA repair (Saredi *et al.* 2016). Taken together, this would situate H4K20me0/1 and PR-Set7 as cell cycle sensor mechanism, with H4K20me0 marking post-replicative chromatin and H4K20me1 identifying G1 chromatin, primed for replication. This scenario would explain both the re-replication phenotype when PR-Set7 is stabilized – aberrant H4K20me1 in G2 mimics G1 cell cycle stage prone for replication – and cell cycle arrest in absence of PR-Set7 – as the cell cannot distinguish between pre- and post-replicative chromatin. Consequently, H4K20me1 might serve as cell cycle stage marker, but is not necessarily directly linked to origin licensing and replication initiation.

HETEROCHROMATIN REPLICATION IS IN PART REGULATED THROUGH H4K20ME3

In contrast, ORC binding to H4K20me3 convincingly connects heterochromatin with replication licensing and activation. In genome-wide ChIP experiments, especially ORC was detected at H4K20me3 sites, while Mcm2-7 was present to a lesser extent (Figure 4.31). Still, functional plasmid abundance experiments proved H4K20me3 being necessary for successful Mcm2-7 loading and origin activity. This result also implies the amount of Mcm2-7 detected by coverage analysis being sufficient for origin licensing. It is possible that an excess of ORC is needed to provide sufficient Mcm2-7 loading at inaccessible chromatin sites. Alternatively, the direct binding capacity of ORC (through the Orc1 BAH domain, Kuo *et al.* 2012; Beck *et al.* 2012) might reduce ORC on-and-off rates and stabilize ORC enough for Mcm2-7 recruitment. Furthermore, ORC might still execute heterochromatin organization functions besides origin licensing. Indeed, ORC has been shown bind heterochromatic structures in yeast, *Xenopus*, *Drosophila*, and mammals. ORC is directly interacting with heterochromatic protein 1 (HP1) and depletion of either protein reduces chromatin binding capacity of the other. In mammals, ORC also interacts with ORC-associated ORCA/LRWD1, which likely facilitates HP1 recruitment to heterochromatin (Chakraborty, Shen, and Prasanth 2011). Furthermore, experiments combining ChIP-seq and mass spectrometry have shown ORCA/ORC complex to directly bind the most prominent repressive histone modifications H3K9me3, H3K27me3 and H4K20me3 (Vermeulen *et al.* 2010). Thereby, it has been proposed that ORC/ORCA specifically recruits lysine methyltransferases and are thus necessary for heterochromatin establishment and maintenance (Giri *et al.* 2015). Furthermore, ORC/ORCA/HP1 was suggested to facilitate origin licensing in heterochromatin (Leatherwood and Vas 2003; Shen *et al.* 2012), however ORC-heterochromatin interaction has already been shown to be separated from licensing functions of ORC (Dillin and Rine 1997; Leatherwood and Vas 2003). My results clearly attribute an origin licensing function to ORC binding to heterochromatic H4K20me3. Genome-wide replication timing experiments thereby revealed that defined late-replicating domains – characterized by the presence of H4K20me3 in unperturbed cells – depend on H4K20me3, while other late-replicating regions did not. These regions might represent heterochromatin where replication is regulated through different mechanisms. In conclusion, H4K20me3-mediated origin licensing and subsequent activation is one mechanism of replication regulation in heterochromatin.

5. CONCLUSION

With this study, I intended to close the gap between extensive analyses of active replication initiation and pre-RC positions. Furthermore, I aimed to elaborate the relation between H4K20 methylation and regulation of DNA replication.

5.1 CHROMATIN ACCESSIBILITY IS THE MAIN DETERMINANT OF PRE-RC POSITIONING IN ACTIVE CHROMATIN

I performed ChIP-seq of the two major pre-RC subcomponents ORC and Mcm2-7, targeting two subunits of each complex (Orc2, Orc3, Mcm3, and Mcm7) to increase validity of the results. Pre-RC ChIP-seq analysis on the EBV genome allowed the determination of the most appropriate peak-calling algorithm, by comparing already published pre-RC positions on the EBV genome with detected pre-RC positions from ChIP-seq. Despite differences in techniques and bioinformatical analyses, the majority of T-PIC-defined pre-RCs detected by ChIP-seq coincided with previously determined pre-RC positions (Papior *et al.* 2012). Comparing replicates of the same ChIP between each other and the different target pre-RC proteins, relatively little variances were observed. ORC, Mcm2-7 and pre-RC positions were concordant and even changed only marginally when ChIP-seq was performed in S/G2 cell cycle stage. This suggests little dynamics of the chromatinized EBV genome, resulting in delimited ORC, Mcm2-7, and pre-RC positions.

Applying the peak-calling settings determined on the EBV genome on the human genome, variances between replicates and between the two pre-RC subcomponents ORC and Mcm2-7 increased. This presumably relies on the definition of a peak as a specific accumulation of reads at a defined position, while neighboring regions decrease in read numbers. However, ORC and Mcm2-7 ChIPs do not meet these requirements. ORC binding has been shown to depend mostly on chromatin accessibility, which renders ORC positions extremely flexible and variable from cell to cell. Consequently, although the region of ORC binding might accord, the precise ORC position fluctuates, which appears as a broad profile from bulk cell ChIP-seq. This impedes the definite detection by a peak calling program, resulting in either non-concordant peaks, or in undetected enrichments due to high read densities, also in neighboring regions. For Mcm2-7, this observation is even more prominent, as Mcm2-7 translocate after loading, rendering peak-detection impossible. To avoid this bias introduced by peak-calling, I additionally relied on the mean sequencing read coverage of each pre-RC component within a specific region of interest. Analyzing the mean read coverage is independent of any peak-calling algorithm and allows a global impression of the average situation at all regions of interest, simultaneously.

Evaluating pre-RC component coverage at active initiation zones revealed pre-RC enrichment within initiation zones, additionally confirming successful pre-RC ChIP-seq. For further analyses, it will be interesting to distinguish between initiation zones flanked or not by actively transcribed genes. As especially Mcm2-7 positioning might depend on transcriptional activity, fundamental differences are expected. Interestingly, pre-RC components, notably Mcm2-7, were specifically depleted from replication termination zones in G1 phase. This observation was cell cycle dependent, as depletion was less prominent in S/G2. Replication termination zones are to a large extent comprised of active genes, consistent with the idea of Mcm2-7 helicase translocation by active transcription machineries.

Generally, pre-RC component coverages were enriched at sites of active transcriptional regulation, like TSS or H3K4me3 sites, in accordance with previous studies. This enrichment was enhanced at an early stage of pre-RC loading (corresponding to S/G2 phase in this study, probably representing a later stage of the cell cycle; needs to be confirmed), implying that within active chromatin, ORC/ pre-RC is first bound to accessible chromatin sites, while Mcm2-7 helicases relocate from their initial loading site prior to S phase (G1 cell cycle stage). Consequently, this study also touches genome-wide pre-RC cell cycle dynamics. The excess of Mcm2-7 compared to ORC situates Mcm2-7 as major determinant of replication initiation.

Employing both HOMER and T-PIC peak-calling, observations from coverage analysis were actually confirmed. Very strong coverage enrichments, as observed at active TSSs or H3K4me3, coincide with strong, HOMER-defined peaks. Active initiation zones were moderately enriched in coverage, however, also sensitive T-PIC-defined peaks were preferably detected in replication initiation zones, when compared to replication termination zones. Consequently, although peak-calling might be misleading for ORC and Mcm2-7 ChIPs, final conclusions are not significantly altered.

My original intention was to unite discrepancies in genome-wide studies of active replication initiation by providing a comprehensive picture of the regulation of pre-RC positioning. In Raji cells, there are no SNS-seq data available, rendering a direct comparison of pre-RC ChIP-seq, SNS-seq and OK-seq impossible. Furthermore, the pre-RC ChIP-seq analysis in hES cells is too preliminary to draw any final conclusions from direct comparisons with SNS-seq. Still, several conclusions emerge from this study:

The major difference between SNS-seq and OK-seq is SNS-seq detecting single replication initiation events, while OK-seq detects zones of preferential initiation, encompassing multiple inefficient initiation sites. This technique disregards single replication origins and correlates with pre-RC coverage analyses, which also do not consider single pre-RC positions. Enrichment of ORC and Mcm2-7 within initiation zones and depletion of these proteins from replication termination zones provide an explanation for the occurrence of preferential replication initiation or termination, depending solely on the density of pre-RC proteins, especially Mcm2-7, on DNA. Only 20% of SNS overlap with initiation zones and vice versa (Petryk *et al.* 2016). However, replication initiation events also occur outside of initiation zones and amount in total to 30000-50000 initiation events per cell. Mapping of more than 200000 replication initiation sites by SNS-seq (Besnard *et al.* 2012) presumably originates from the large distribution of Mcm2-7, resulting in many different initiation sites in a bulk cell population. It might be possible, that the 200000 detected origins are the detection limit of the SNS-seq technique, by employing a limited number of cells and a certain size selection of nascent DNA. It might be discussable whether many more replication initiation sites could possibly be detected encompassing all sites of all cells.

Single-cell replication initiation experiments might provide answers for that question. Combination of various single-cell experiments would reveal whether the final number of detected initiation events evens out around 200000 events or continues to grow exponentially. Also single-cell pre-RC ChIPs would help to better understand the context. However, although single-cell experiments are advancing, both single-cell SNS-seq or pre-RC ChIP-seq are not feasible to date. Currently broad pre-RC profiles originating from bulk ChIP-seq experiments complicate conclusions about precise ORC or Mcm2-7 positions. Knowing the positions in a single cell and extrapolating the results would indeed complete the idea of variable ORC and Mcm2-7 positioning.

In conclusion, this study provided new insights in the regulation of pre-RC positioning and implies Mcm2-7 being the main determinant of DNA replication in accessible chromatin regions characterized by transcriptional activity.

5.2 DIRECT ORC-CHROMATIN INTERACTIONS REGULATE PRE-RC POSITIONING IN HETEROCHROMATIN

H4K20 methylation was also assessed in this study, to decipher the role of this histone modification in replication regulation. Altering the expression of the Histone 4 Lysine 20 monomethyltransferase PR-Set7 has severe DNA replication phenotypes. Consequently, H4K20me1 was expected to be involved in the regulation of replication licensing and/or activation. However, pre-RC displayed no striking association to H4K20me1 in euchromatin in ChIP-seq experiments. This result situates PR-Set7 being a major regulator of the cell cycle rather than DNA replication mechanisms themselves. Interestingly, especially ORC associated with H4K20me3 in heterochromatin, while binding of Mcm2-7 was less prominent. Because of reported roles of ORC in heterochromatin organization, I functionally tested implication of H4K20me3 in origin licensing and confirmed that H4K20me3 is necessary for Mcm2-7 recruitment and subsequent replication activation, in specific heterochromatin domains.

Consequently, euchromatin and heterochromatin differ in ORC loading mechanisms. While in euchromatin, ORC randomly binds accessible chromatin sites and loads multiple Mcm2-7 complexes, in heterochromatin ORC-binding is stabilized by specific histone modifications. In my study, H4K20me3 regulates replication in specific heterochromatin regions. H4K20me3 is necessary for pre-RC formation within these regions and also controls pre-RC activation. Combining my observations and current literature led to the proposition of the following model (Figure 5.1):

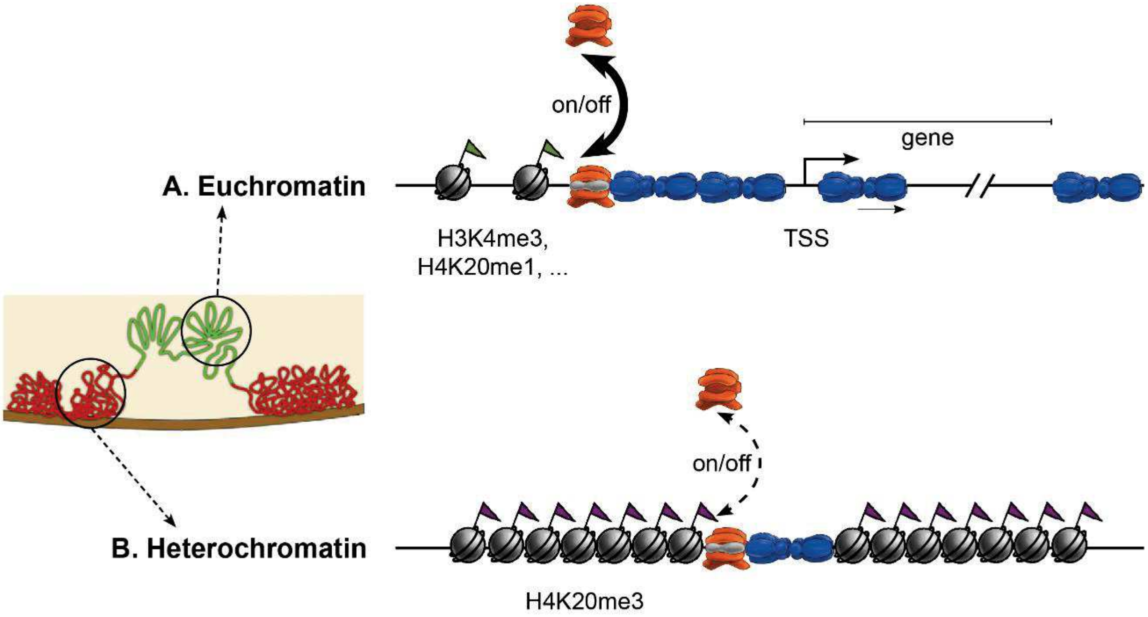


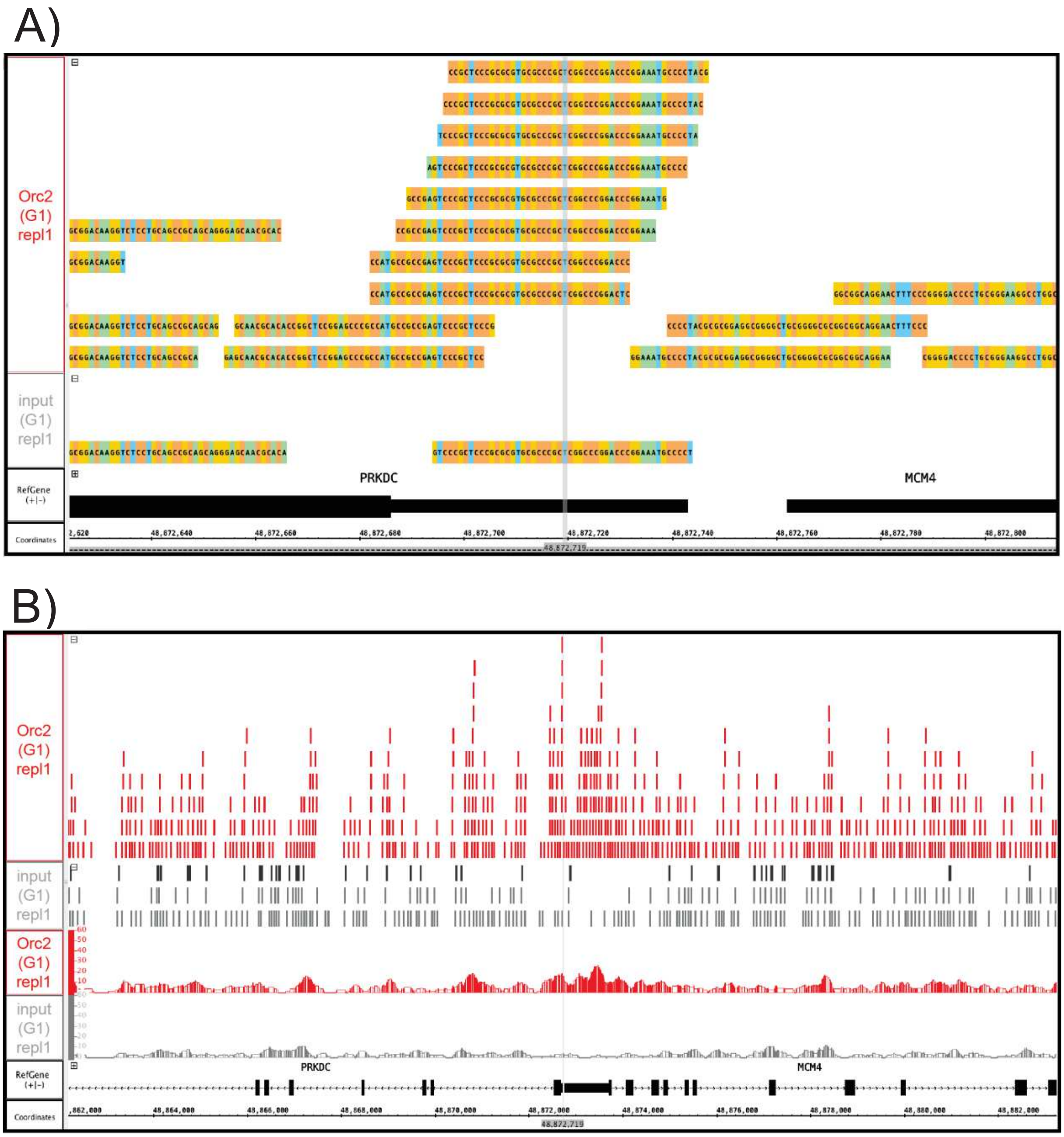
FIGURE 5.1: MODEL FOR REPLICATION LICENSING MECHANISMS IN EUCHROMATIN AND HETEROCHROMATIN. A) OPEN CHROMATIN STRUCTURES AND ACTIVE TRANSCRIPTION MEDIATE MCM2-7 LOADING AND TRANSLOCATION. High ORC on- and off-rates lead to excessive Mcm2-7 loading. Active transcription machineries translocate Mcm2-7 helicases to gene-adjacent regions. B) IN HETEROCHROMATIN, H4K20ME3 STABILIZES ORC BINDING. Lower on- and off-rates lead to less Mcm2-7 recruitment.

A. Euchromatin locates interior of the nucleus, is marked by active chromatin marks (like H3K4me3) and is actively transcribed. ORC favors accessible chromatin regions (like TSS) and binds these regions. High on- and off-rates of ORC or chromatin conformation lead to multiple Mcm2-7 hexamer loading, which are translocated by the transcriptional machinery to gene-adjacent regions. Elevated Mcm2-7 densities increase activation probability in S phase and lead to early origin firing.

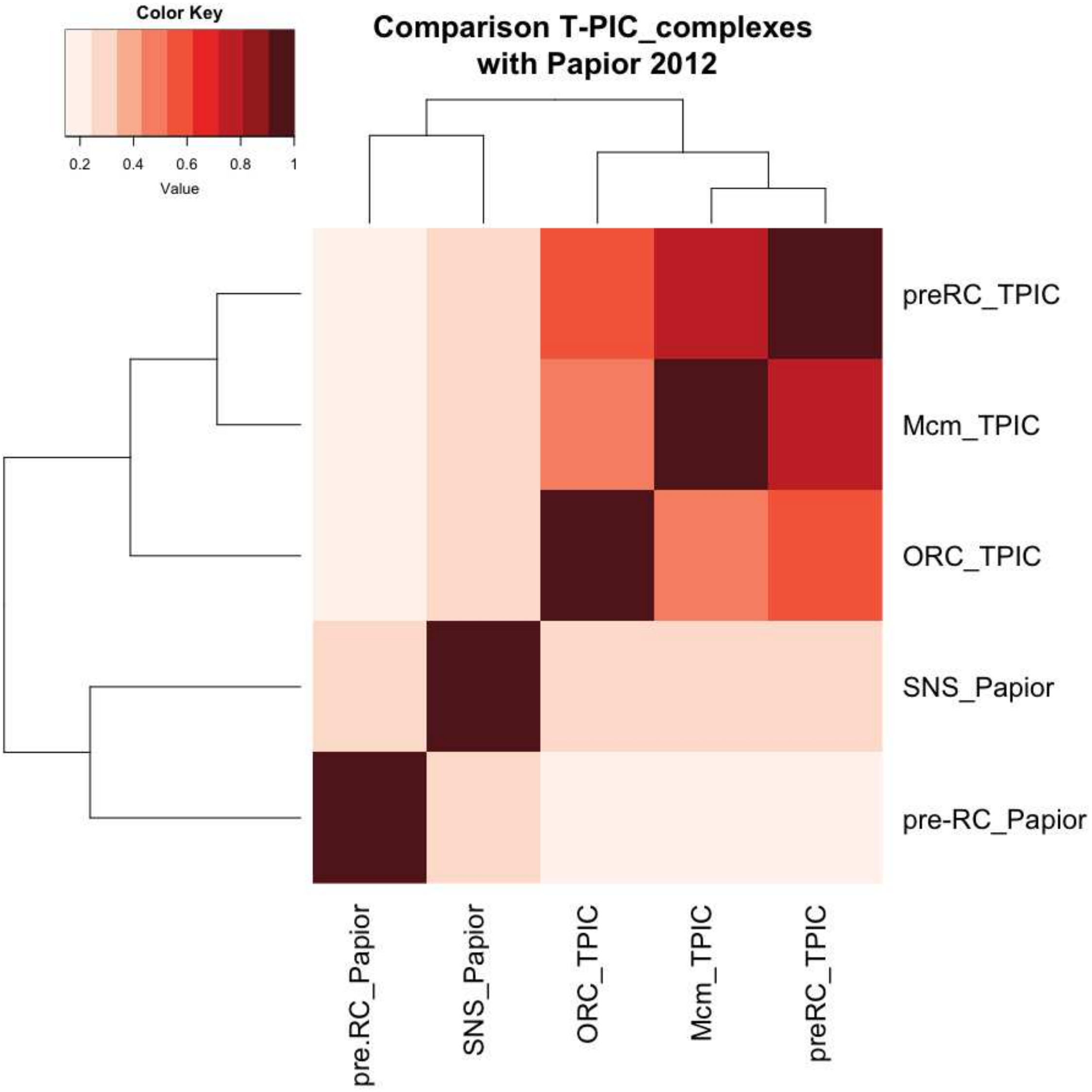
B. Heterochromatin locates closer to the nuclear lamina. ORC binding is impaired through less accessible chromatin structure. Consequently, direct interactions of histone modifications (like H4K20me3) and ORC mediate ORC association to DNA. Stabilization of ORC leads to lower on- and off-rates and less Mcm2-7 recruitment. However, one Mcm2-7 double-hexamer is theoretically enough for replication initiation although reduced firing probabilities situate heterochromatin replication in late S phase.

Taken together, this work constitutes the first genome-wide ChIP-seq analysis of a full set of pre-RC components, as especially Mcm2-7 has not been analyzed so far. It situates Mcm2-7 as major determinant of replication initiation in euchromatin, while ORC displays the major regulator in heterochromatin. Furthermore, it provided direct evidence of replication origin licensing in heterochromatin depending on H4K20me3. These results added a piece to the puzzle of human replication regulation. Some already existing pieces might require re-evaluation, while future investigations will certainly provide other missing elements to complete the picture.

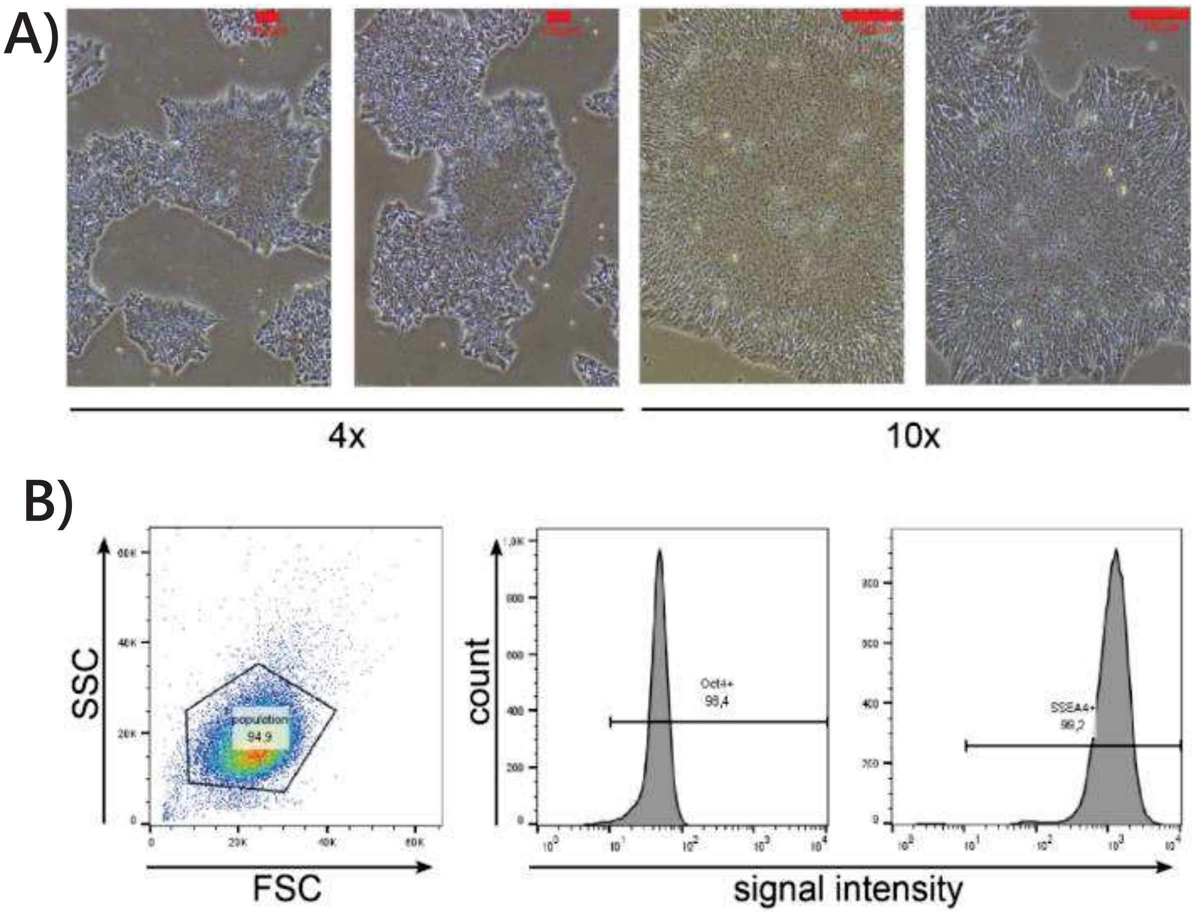
APPENDIX



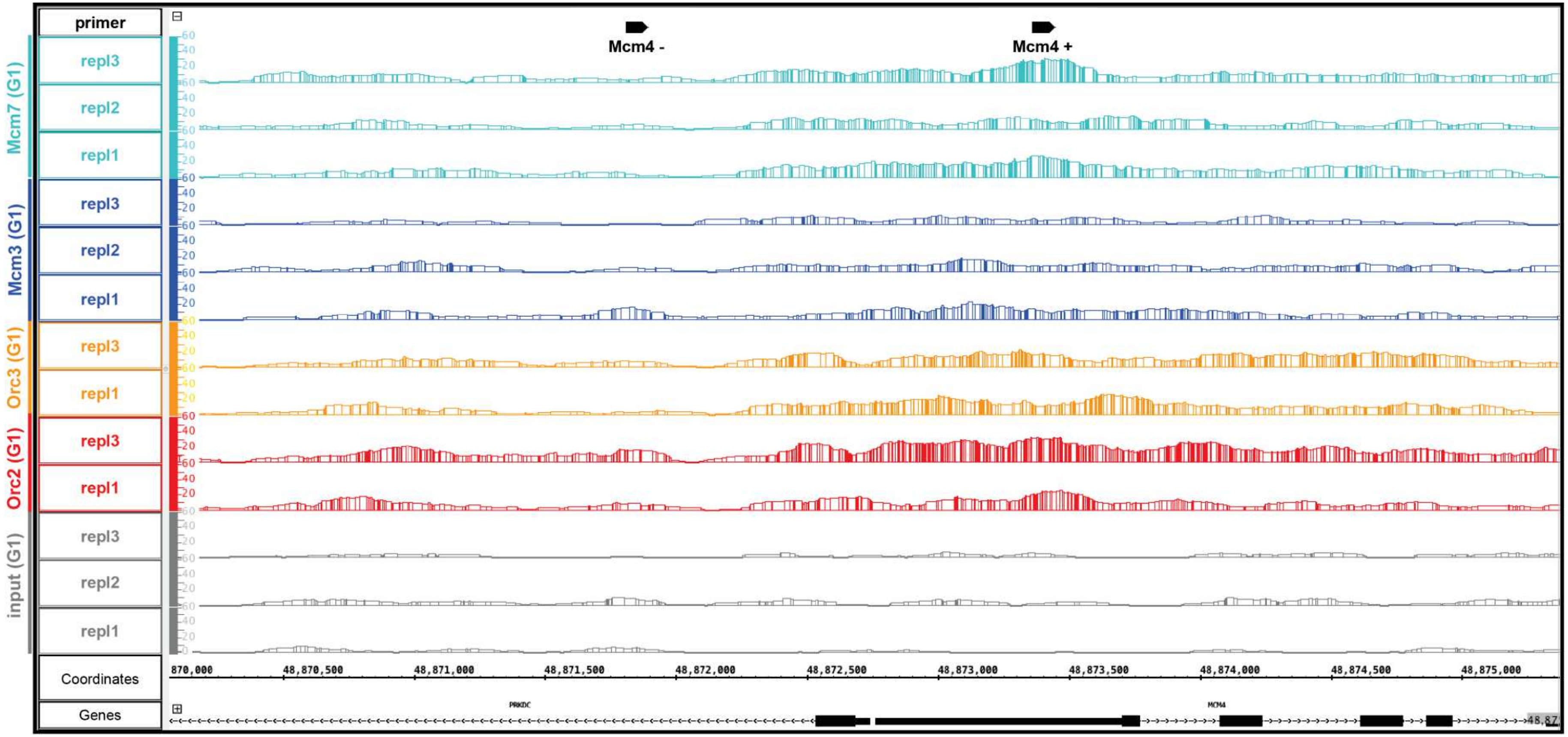
APPENDIX FIGURE 1: DIFFERENT REPRESENTATIONS OF INPUT (GREY) AND ORC2 (RED) CHIP-SEQ SAMPLES IN IGB AT THE MCM4/PRKDC LOCUS. A) SEQUENCING ALIGNMENT AFTER MAPPING AGAINST THE HUMAN GENOME. UTRs of divergent Mcm4 and PRKDC genes are visible as black bars, the first PRKDC exon as thicker black bar. [chr8: 48.872.624-48.872.814]. B) DIRECT COMPARISON OF SEQUENCING ALIGNMENT AND PROFILE REPRESENTATION. Upper two panels: Zoom-out of sequencing alignment from A); Lower two panels: profile representation of the same samples. introns = thin lines, arrowheads point to direction of transcription. Thicker black bars = exons, thin black bars = UTR. [chr8: 48.862.188-48.883.250].



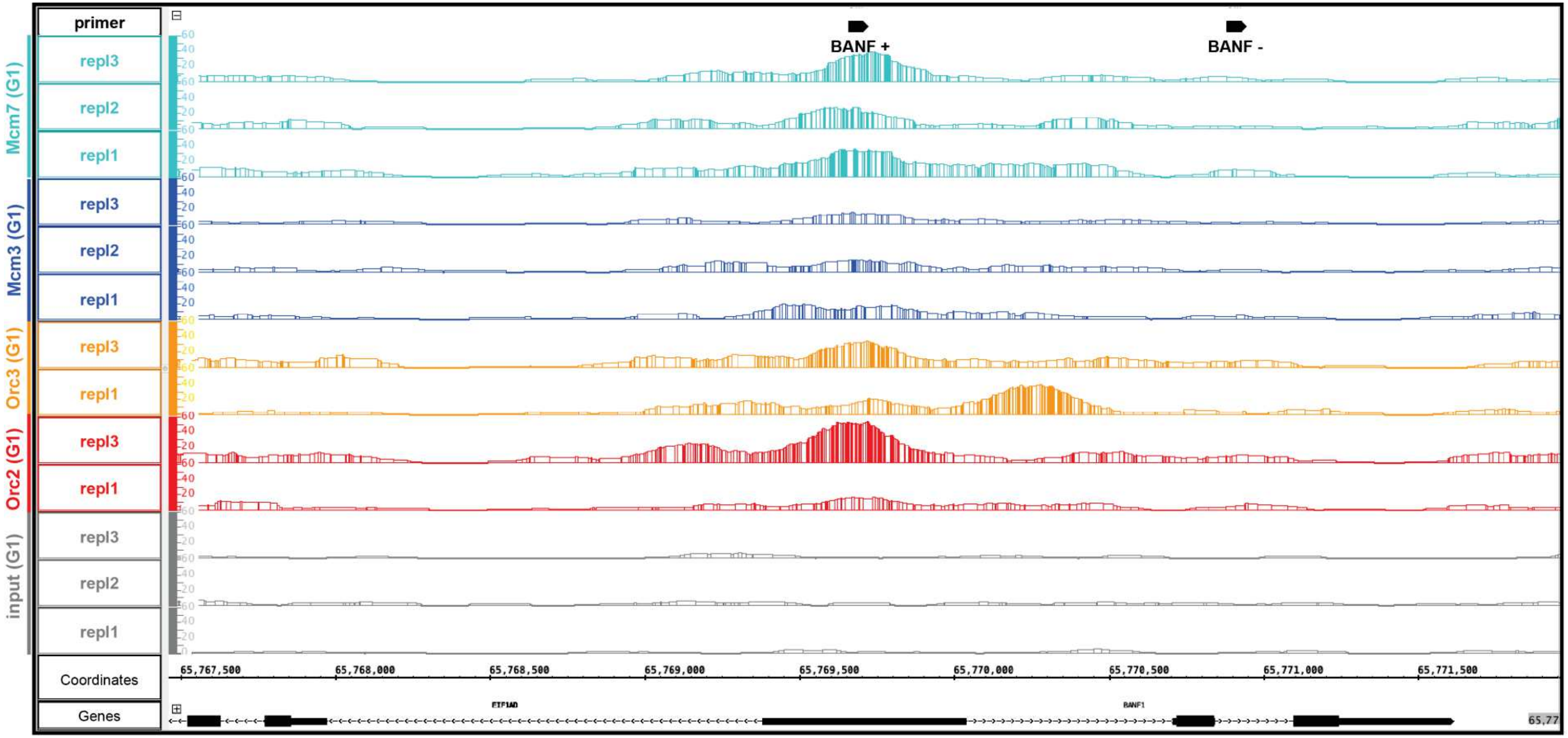
APPENDIX FIGURE 2: HEATMAP REPRESENTATION OF THE JACCARD SIMILARITY INDEX OF T-PIC-DEFINED COMPLEXES AND PRE-RC (PAPIOR *ET AL.* 2012) AND SNS (PAPIOR *ET AL.* 2012). From little similarity (light red) to high similarity (dark red).



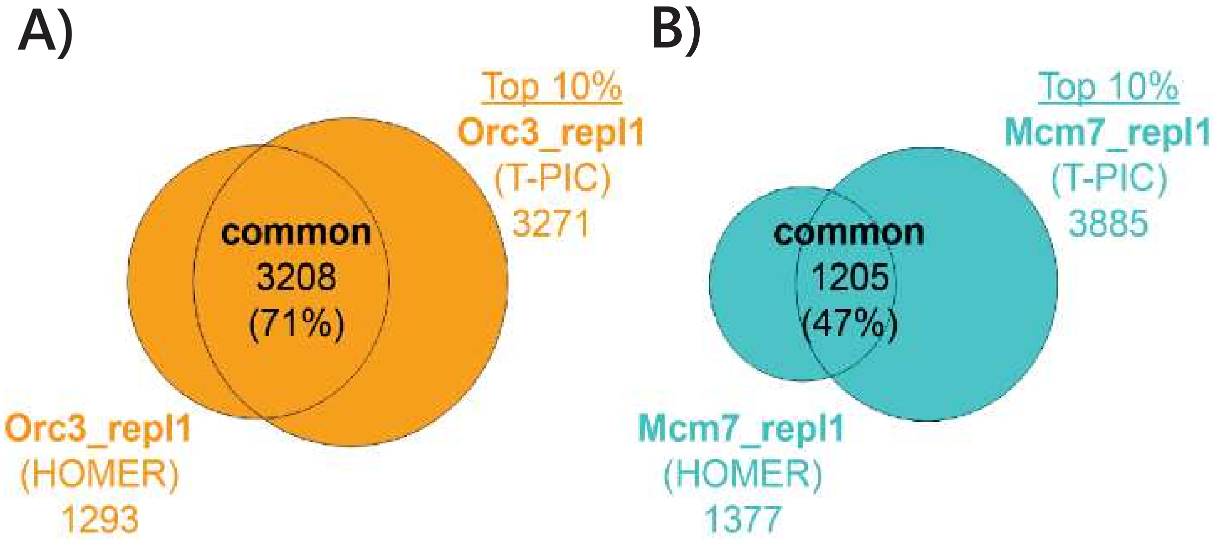
APPENDIX FIGURE 3: VALIDATION OF ES CELL PLURIPOTENCY. A) CELL MORPHOLOGY. Red bar represents 100µm. B) FACS STAIN OF PLURIPOTENCY MARKERS OCT4 AND SSEA4. Staining was performed according to BD Stemflow Human and Mouse Pluripotent Stem Cell Analysis Kit. FACS was calibrated using the corresponding isotype negative control.



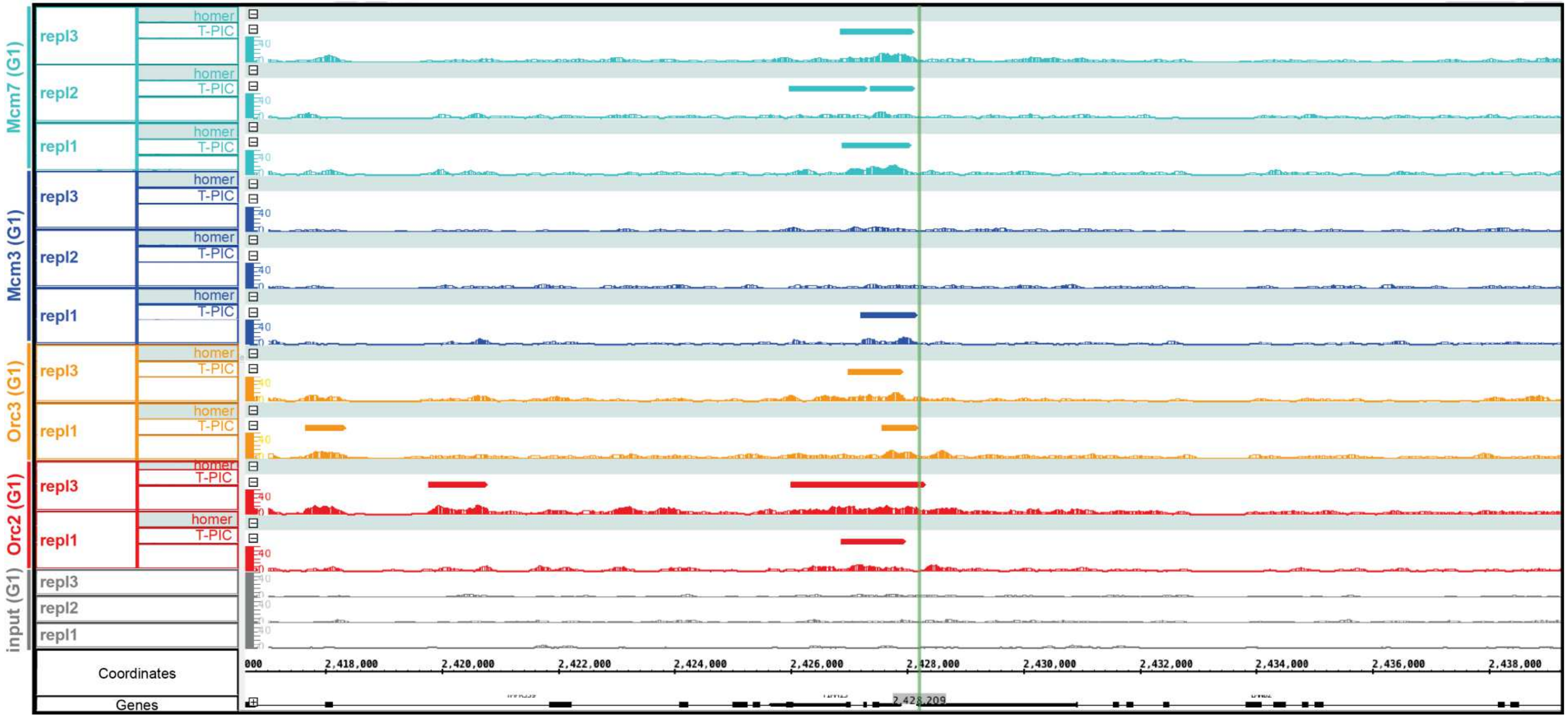
APPENDIX FIGURE 4: IGB VISUALIZATION OF THE MCM4/PRKDC LOCUS FOR PRIMER DEFINITION. Primer positions are indicated as black bars on top. Positive primer marked by Mcm4+, negative primer marked Mcm4-. Input: grey, Orc2: red, Orc3: orange, Mcm3: blue, Mcm7: turquois. [chr8: 48870064-48875373].



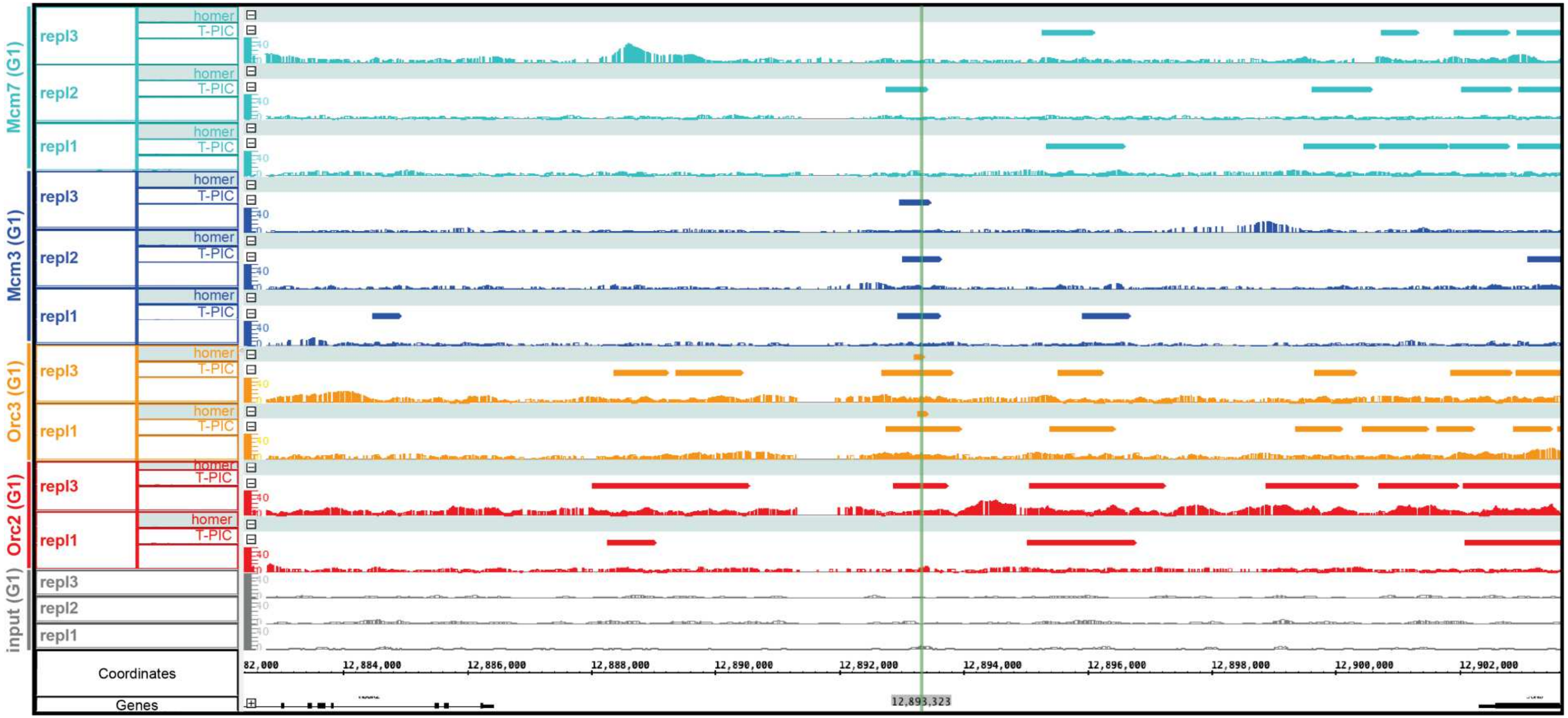
APPENDIX FIGURE 5: IGB VISUALIZATION OF THE BANF LOCUS FOR PRIMER DEFINITION. Primer positions are indicated as black bars on top. Positive primer marked by BANF+, negative primer marked BANF-. Input: grey, Orc2: red, Orc3: orange, Mcm3: blue, Mcm7: lightblue. [chr11: 65767460-65771961]



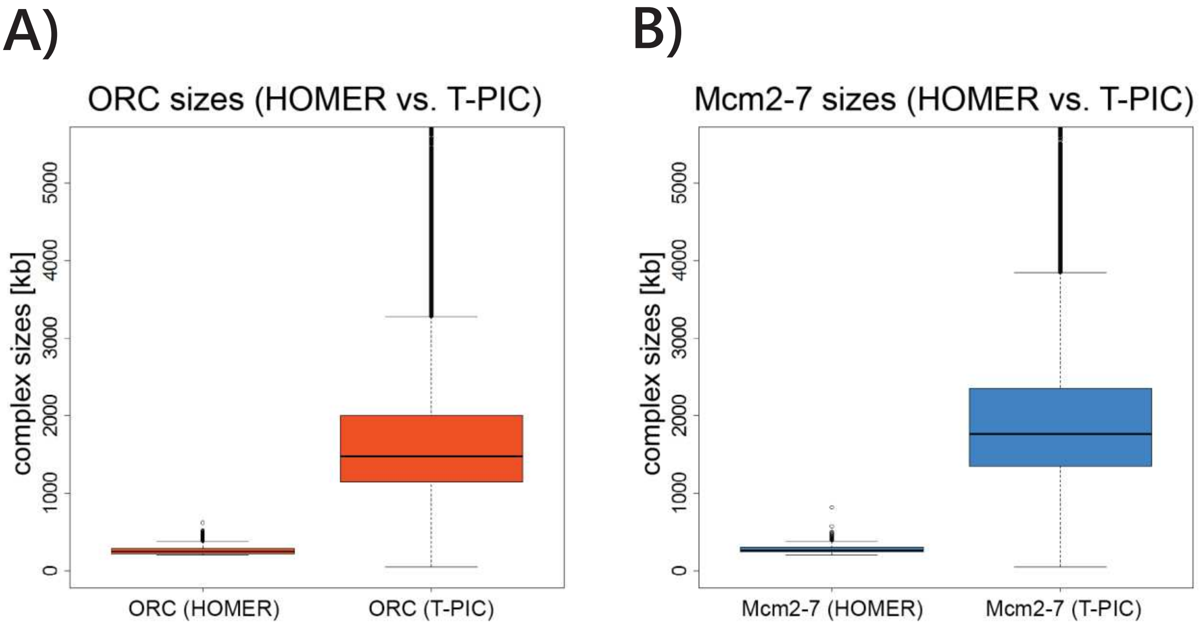
APPENDIX FIGURE 6: MOST HOMER PEAKS RESIDE WITHIN THE TOP 10% OF T-PIC PEAKS. REPRESENTATIVE EXAMPLE ANALYSIS FOR A) ORC3 AND B) MCM7 IN ONE REPLICATE. Venn diagram of overlap between all HOMER-detected peaks with the top 10% of T-PIC-detected peaks. Overall counts are indicated. The percentage of overlapping HOMER-detected peaks are specified in brackets.



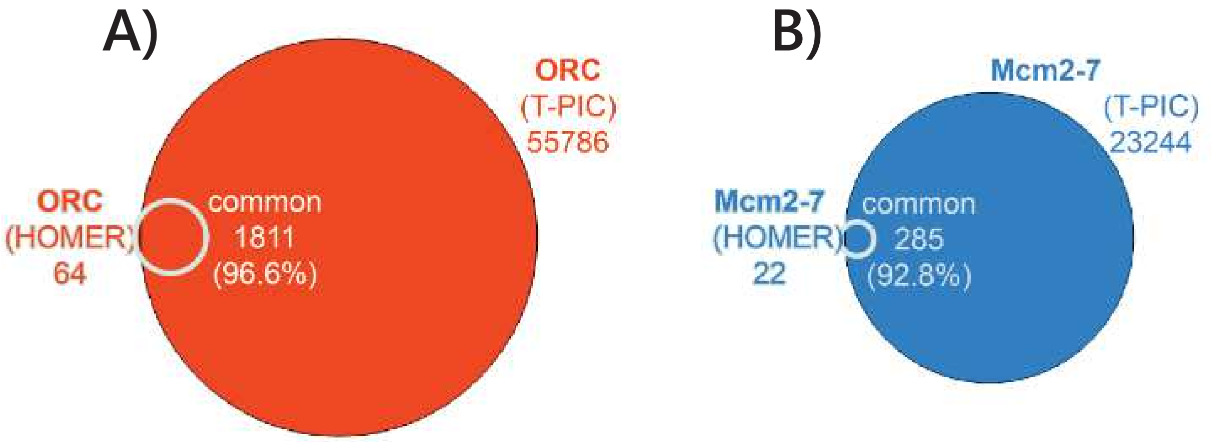
APPENDIX FIGURE 7: PRE-RC BINDING AT LAMINB2 ORIGIN IS DETECTED BY T-PIC BUT NOT BY HOMER. Input: grey, Orc2: red, Orc3: orange, Mcm3: blue, Mcm7: turquois. T-PIC-detected peaks are visualized as bars; HOMER-detected peaks are additionally highlighted by lightblue background. Position of LaminB2 origin is marked by green horizontal line. [chr19: 2416622-2440010].



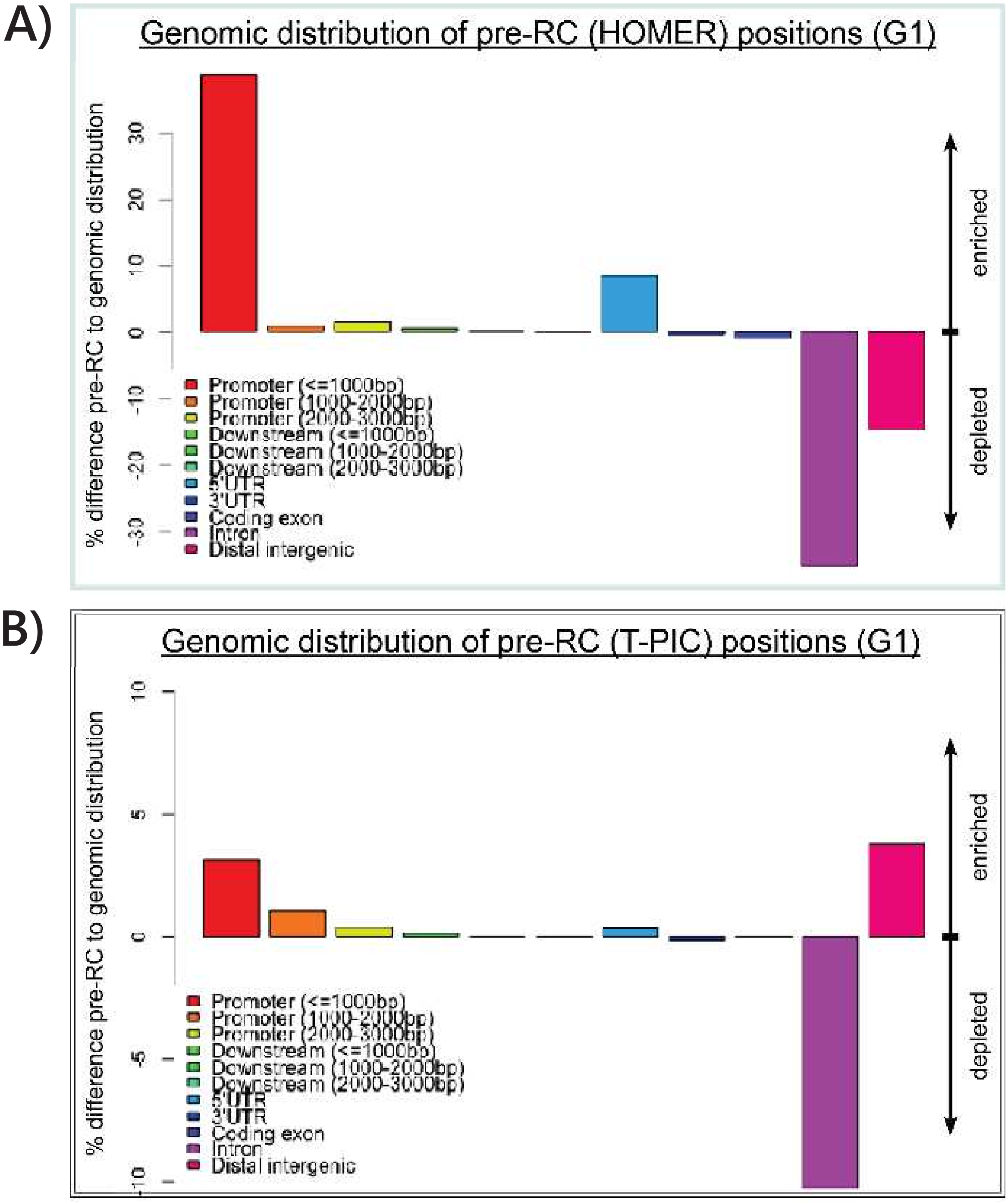
APPENDIX FIGURE 8: PRE-RC BINDING AT JUNB ORIGIN IS DETECTED BY T-PIC BUT NOT BY HOMER. Input: grey, Orc2: red, Orc3: orange, Mcm3: blue, Mcm7: turquois. T-PIC-detected peaks are visualized as bars; HOMER-detected peaks are additionally highlighted by lightblue background. Position of JunB origin is marked by green horizontal line. [chr19: 12882387-12904261].



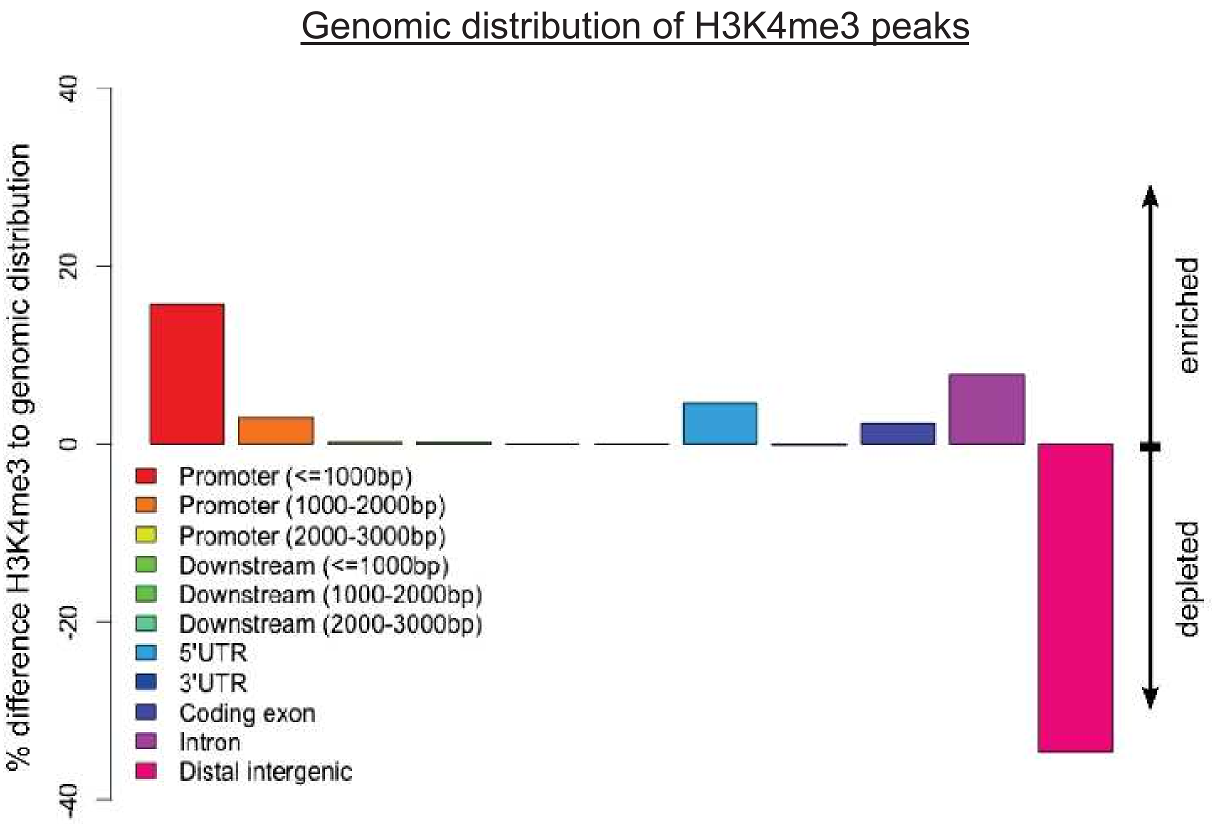
APPENDIX FIGURE 9: ORC AND MCM2-7 COMPLEX SIZES (HOMER VS. T-PIC). A) ORC. B) MCM2-7. Represented in boxplot: thick line shows the median, the box is the distribution from the first to the third quartile, the whiskers indicate the smallest and largest value without being an outlier. Outliers represented by dots.



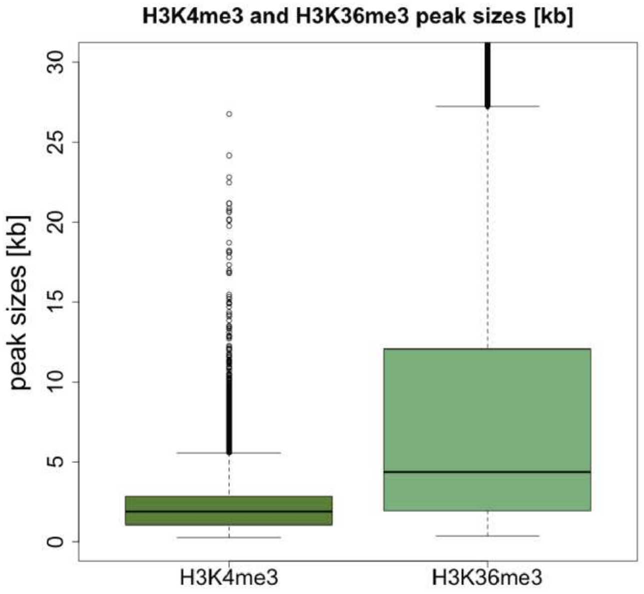
APPENDIX FIGURE 10: OVERLAPS OF HOMER- AND T-PIC DEFINED COMPLEXES. A) ORC B) MCM2-7. Venn diagram of overlap between HOMER- and T-PIC-defined complexes. Overall counts are indicated. The percentage of HOMER-defined complexes common with T-PIC-defined pre-RCs is specified in brackets.



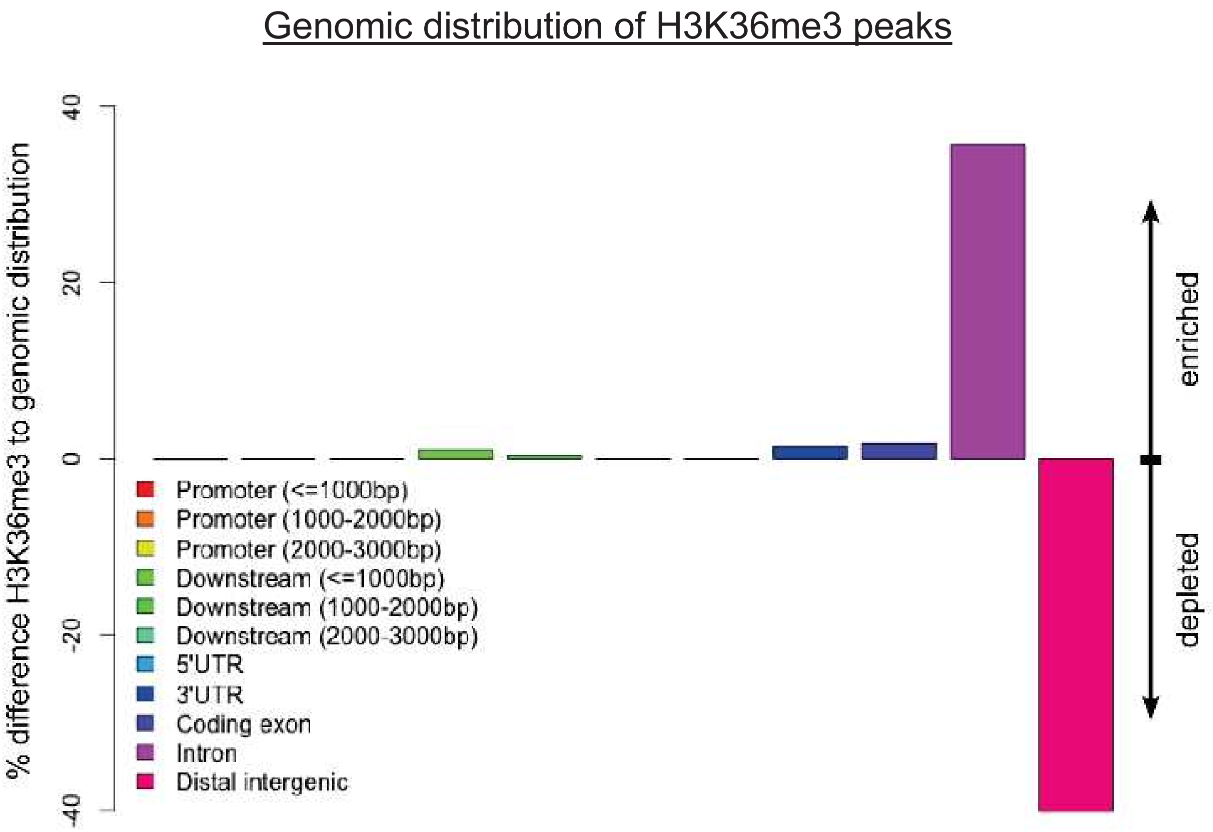
APPENDIX FIGURE 11: GENOMIC DISTRIBUTION OF A) HOMER-DEFINED PRE-RC AND B) T-PIC DEFINED PRE-RC REVEALED ASSOCIATION TO PROXIMAL PROMOTER REGIONS. The genomic distribution was calculated using the CEAS program and normalized against the "default" genomic distribution of the single criteria (position 0 on the y-axis). Distribution was then represented with upwards orienting bars showing % enrichments and downwards oriented bars showing % depletion compared to "default" genomic distribution.



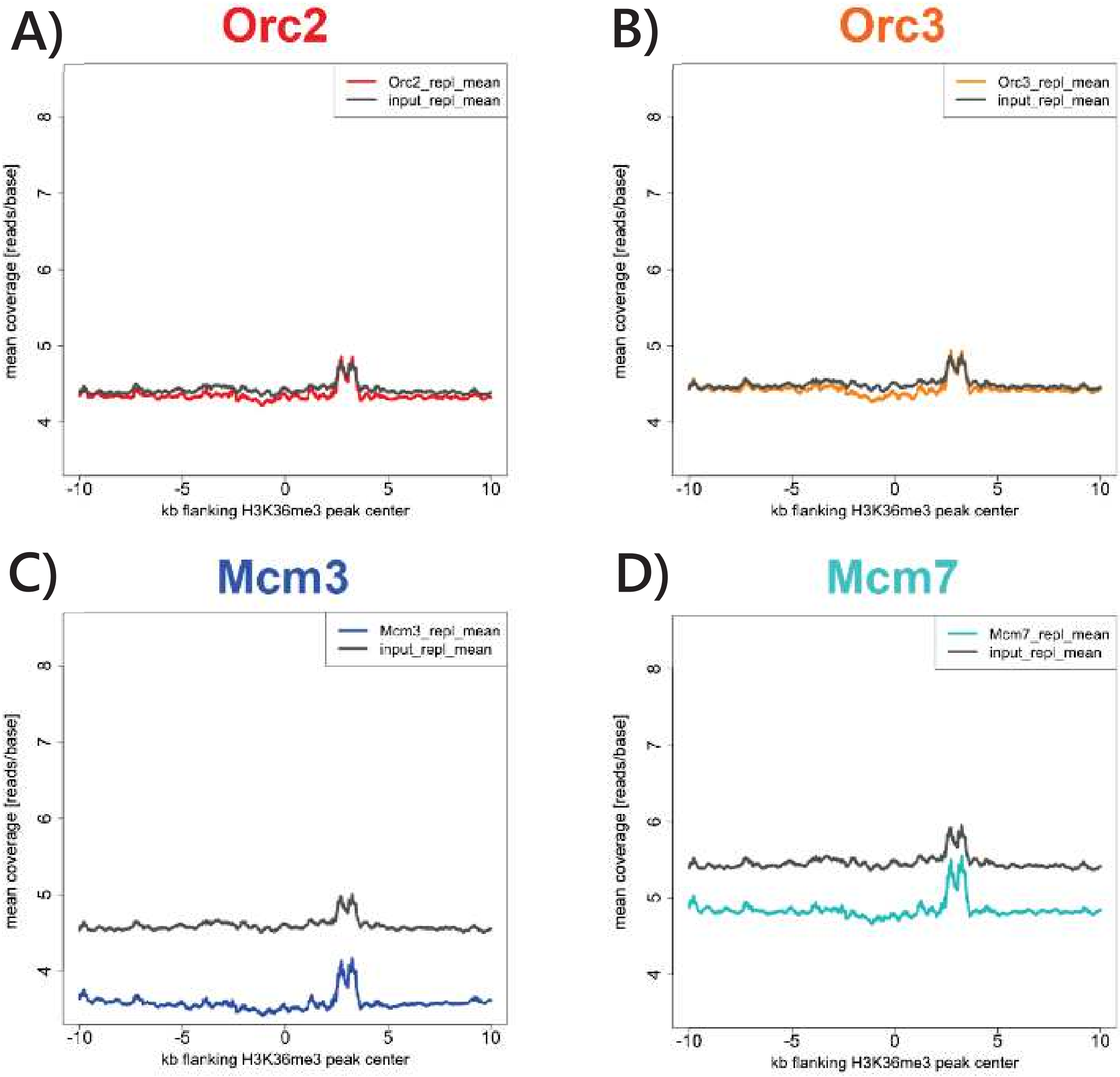
APPENDIX FIGURE 12: GENOMIC DISTRIBUTION OF H3K4ME3 PEAKS SHOWED CLOSE ASSOCIATION TO REGULATORY 5‘ GENIC REGIONS. The genomic distribution was calculated using the CEAS program and normalized against the "default" genomic distribution of the single criteria (position 0 on the y-axis). Distribution was then represented with upwards orienting bars showing % enrichments and downwards oriented bars showing % depletion compared to "default" genomic distribution.



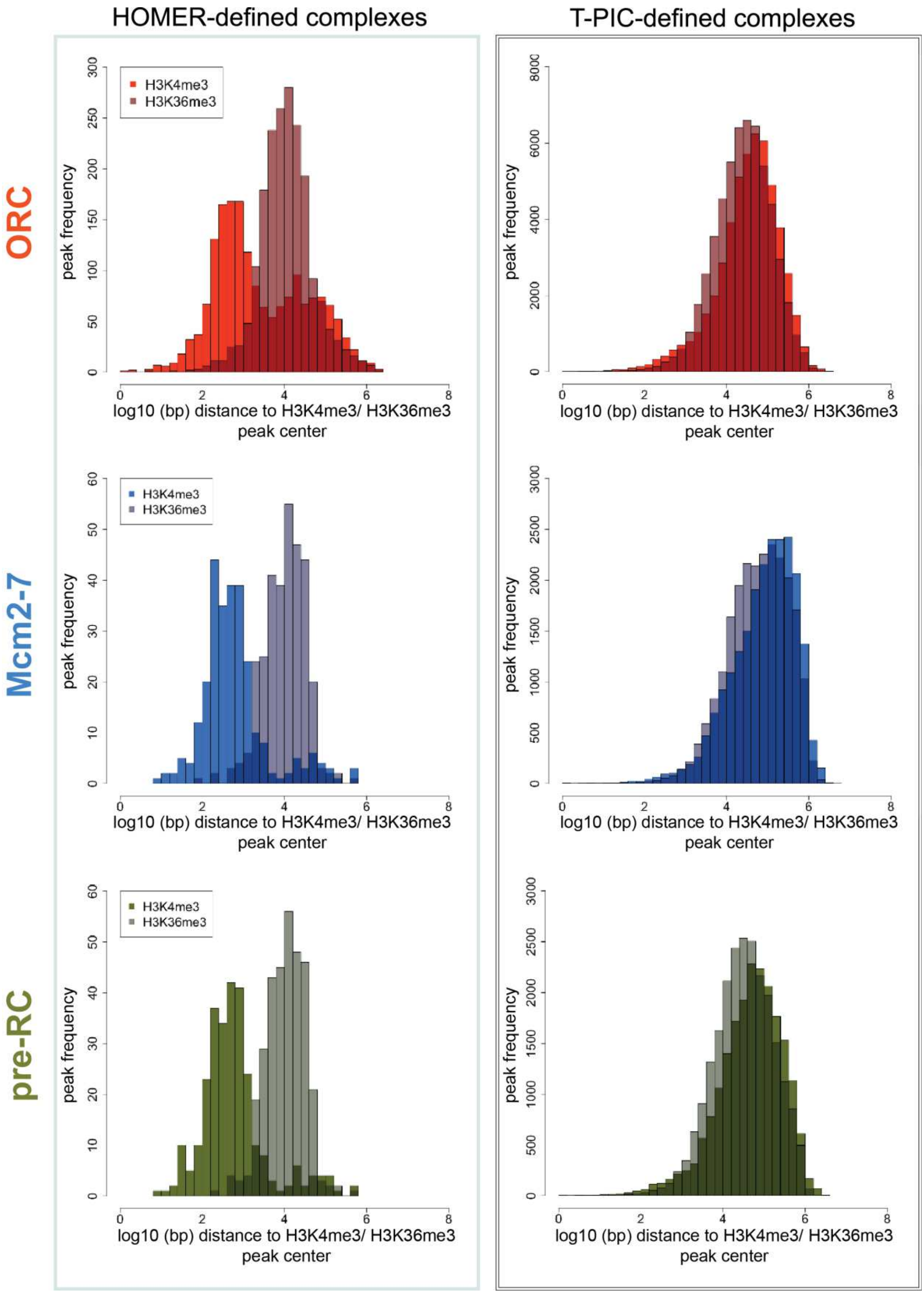
APPENDIX FIGURE 13: H3K4ME3 POSITIONS WERE MORE LOCALIZED THAN H3K36ME3. Peak sizes in [kb] are represented in boxplots: thick line shows the median, the box is the distribution from the first to the third quartile, the whiskers indicate the smallest and largest value without being an outlier. Outliers represented by dots.



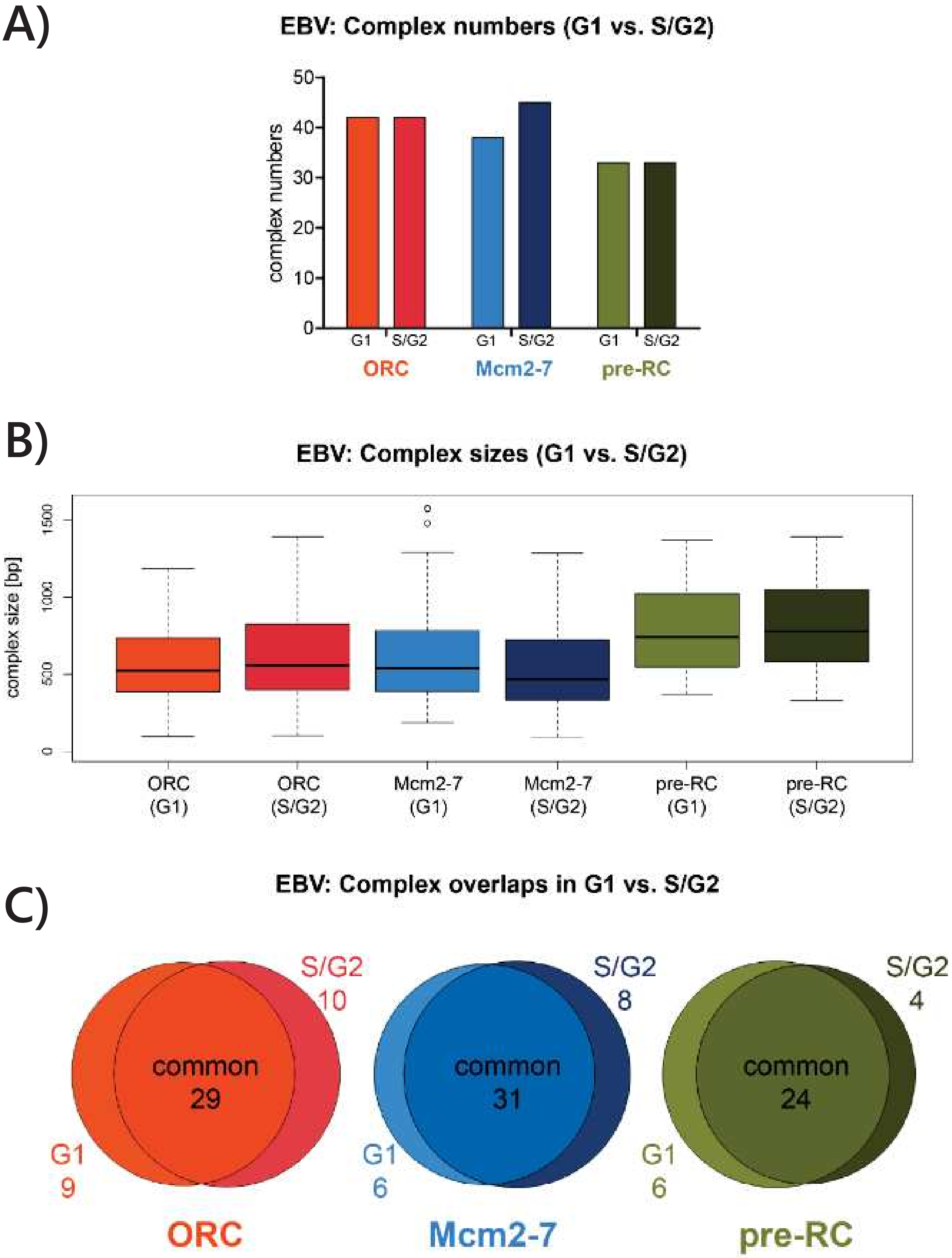
APPENDIX FIGURE 14: GENOMIC DISTRIBUTION OF H3K36ME3 PEAKS SHOWED ENRICHMENT IN INTRONS. The genomic distribution was calculated using the CEAS program and normalized against the "default" genomic distribution of the single criteria (position 0 on the y-axis). Distribution was then represented with upwards orienting bars showing % enrichments and downwards oriented bars showing % depletion compared to "default" genomic distribution.



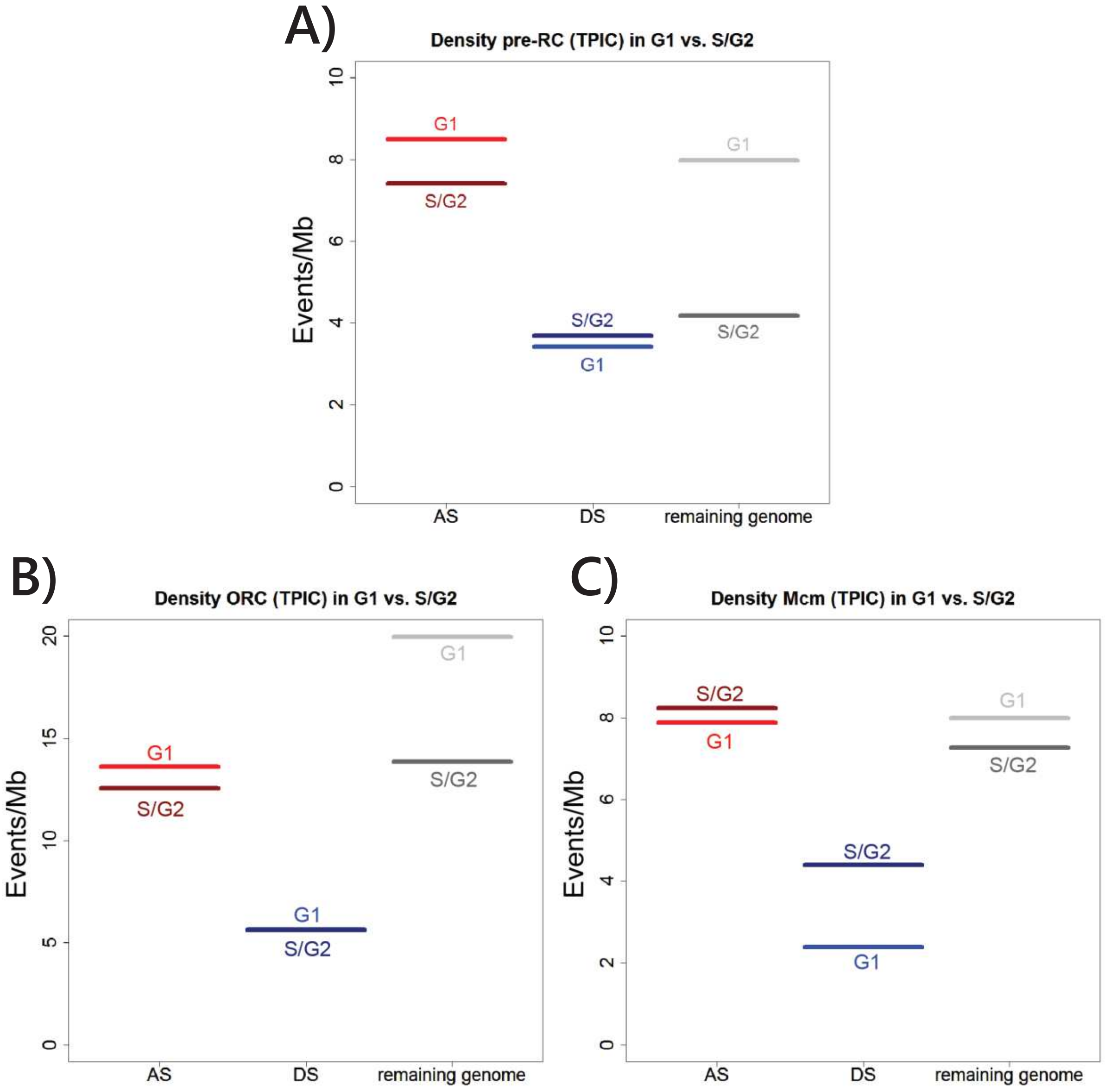
APPENDIX FIGURE 15: NO ENRICHED COVERAGE OF PRE-RC PROTEINS AT H3K36ME3 PEAKS. A) MEAN ORC2 COVERAGE. B) MEAN ORC3 COVERAGE. C) MEAN MCM3 COVERAGE. D) MEAN MCM7 COVERAGE. Coverage was calculated as number of reads/base within a 20 kb window around H3K36me3 peak center. Input was plotted as control (grey).



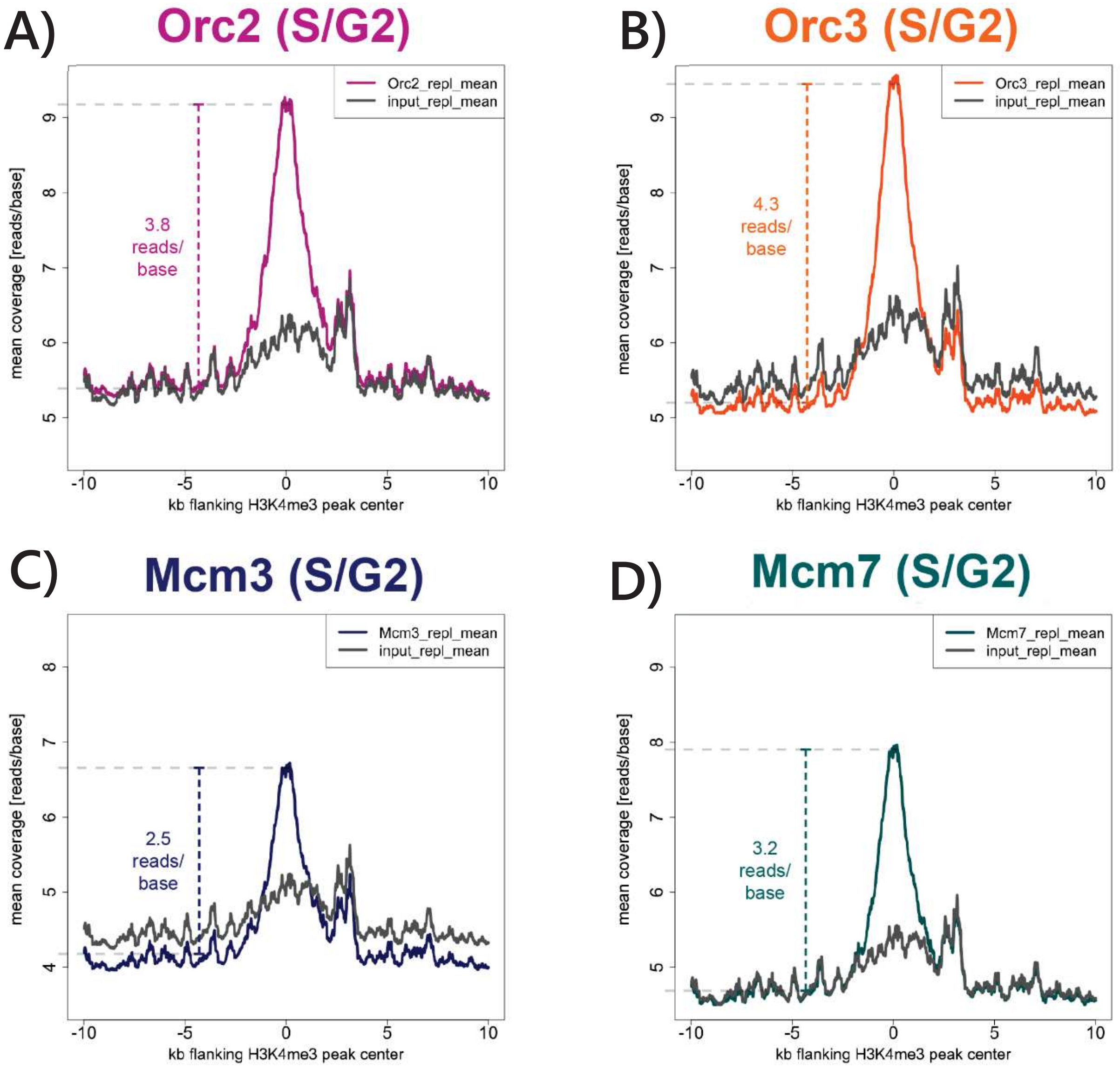
APPENDIX FIGURE 16: HOMER-DEFINED ORC, MCM2-7, AND PRE-RC WERE CLOSELY ASSOCIATED TO H3K4ME3 AND NOT H3K36ME3. Left: HOMER-defined complexes, right: T-PIC-defined complexes. The distance of each position was calculated towards the next H3K4me3 or H3K36me3 peak center and is plotted in log10 on the x-axis. The frequency of complexes within a specific distance is plotted on the y-axis.



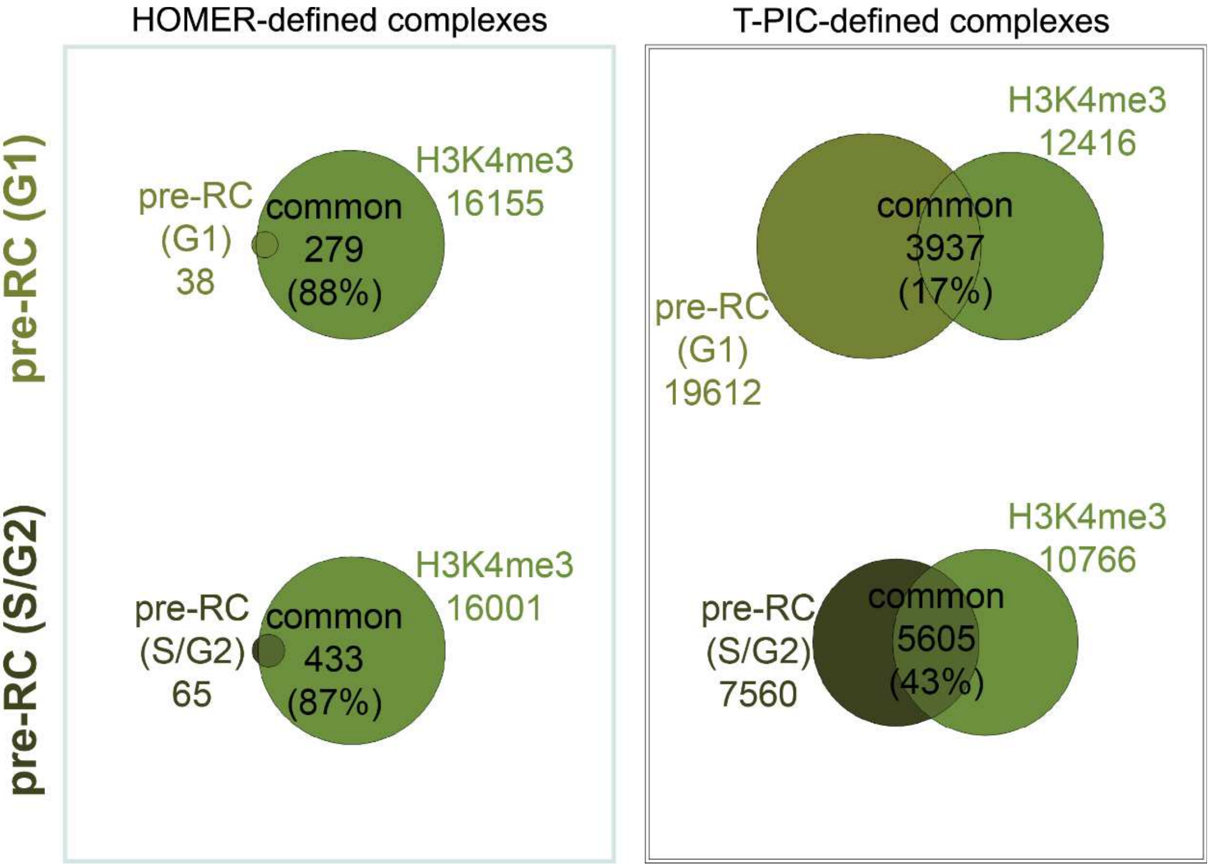
APPENDIX FIGURE 17: T-PIC-DEFINED ORC, MCM2-7, AND PRE-RC ARE VERY SIMILAR IN G1 AND S/G2 ON EBV. A) COMPLEX NUMBERS. Number of T-PIC-defined ORC, Mcm2-7 and pre-RC was plotted as indicated in a bar chart in G1 vs. S/G2. B) COMPLEX SIZES. Complex sizes were potted as boxplot in G1 vs. S/G2. Thick line shows the median, the box is the distribution from the first to the third quartile, the whiskers indicate the smallest and largest value without being an outlier. Outliers represented by dots. C) COMPLEX POSITION OVERLAPS. Venn diagram of overlap between G1-defined ORC, Mcm2-7, or pre-RC and the same complexes defined in S/G2. Overall counts are indicated.



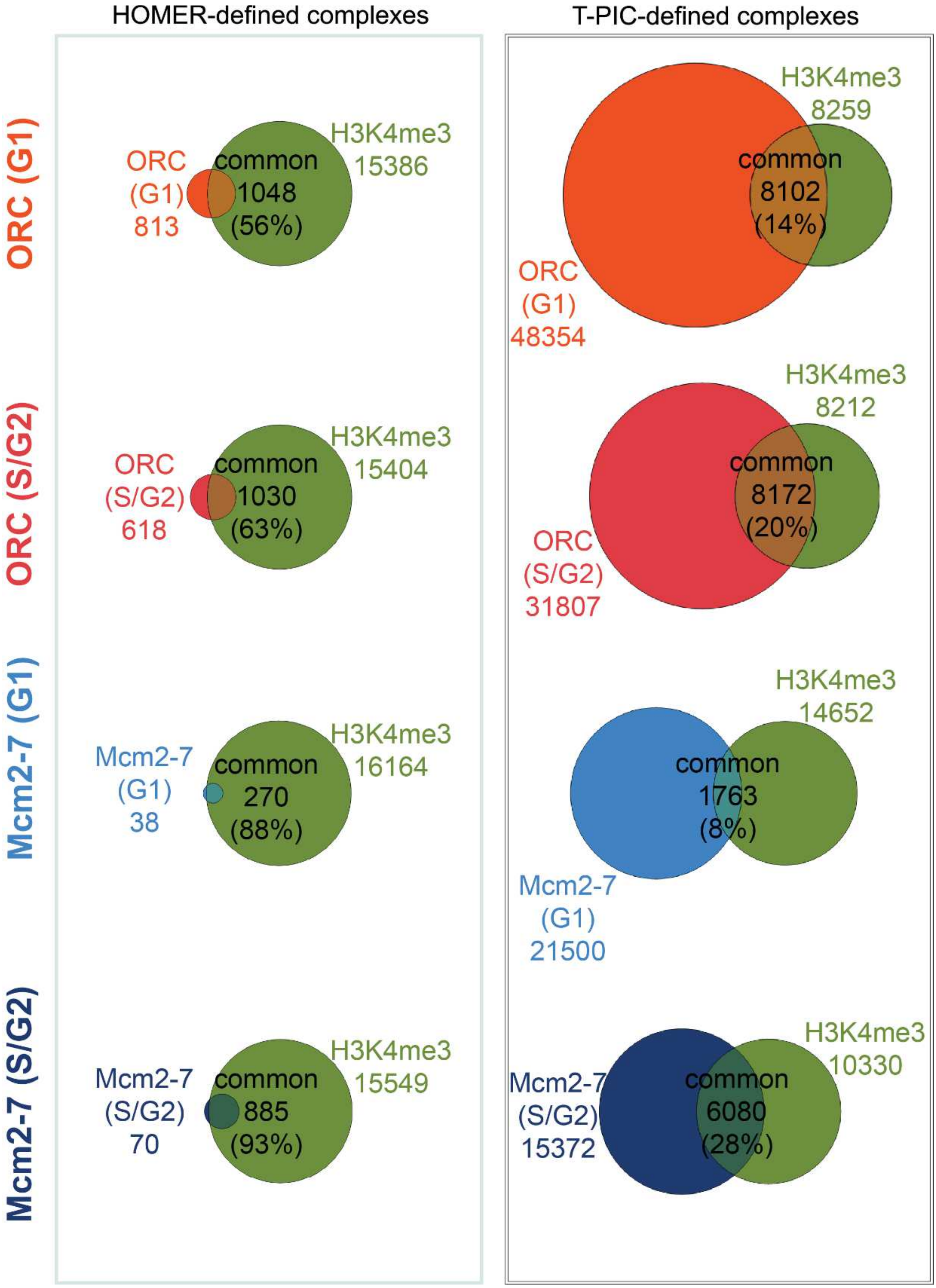
APPENDIX FIGURE 18: COMPARISON OF T-PIC-DEFINED PRE-RC, ORC AND MCM2-7 DENSITIES AT AS, DS, AND THE REMAINING GENOME IN G1 AND S/G2. A) T-PIC-DEFINED PRE-RC DENSITIES. B) T-PIC-DEFINED ORC DENSITIES. C) T-PIC-DEFINED MCM2-7 DENSITIES. The cell cycle stage is represented above/beneath each line, S/G2 densities are represented in darker colors.



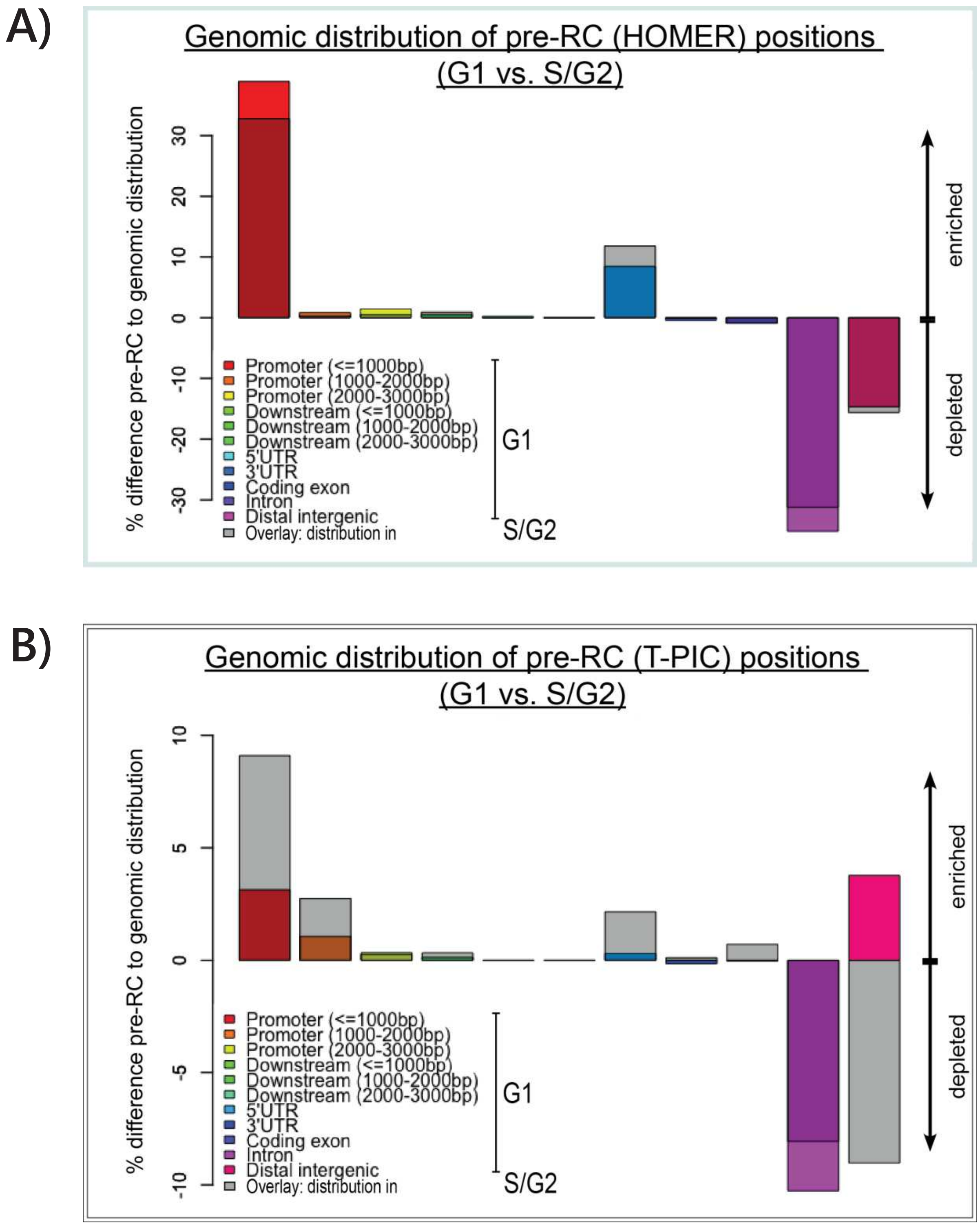
APPENDIX FIGURE 19: PRE-RC PROTEIN COVERAGE IN S/G2 AT H3K4ME3 PEAKS IS MORE PROMINENT THAN IN G1. A) MEAN ORC2 S/G2 COVERAGE. B) MEAN ORC3 S/G2 COVERAGE. C) MEAN MCM3 S/G2 COVERAGE. D) MEAN MCM7 S/G2 COVERAGE. Coverage was calculated as number of reads/base within a 20 kb window around H3K4me3 peak center. Input was plotted as control (grey).



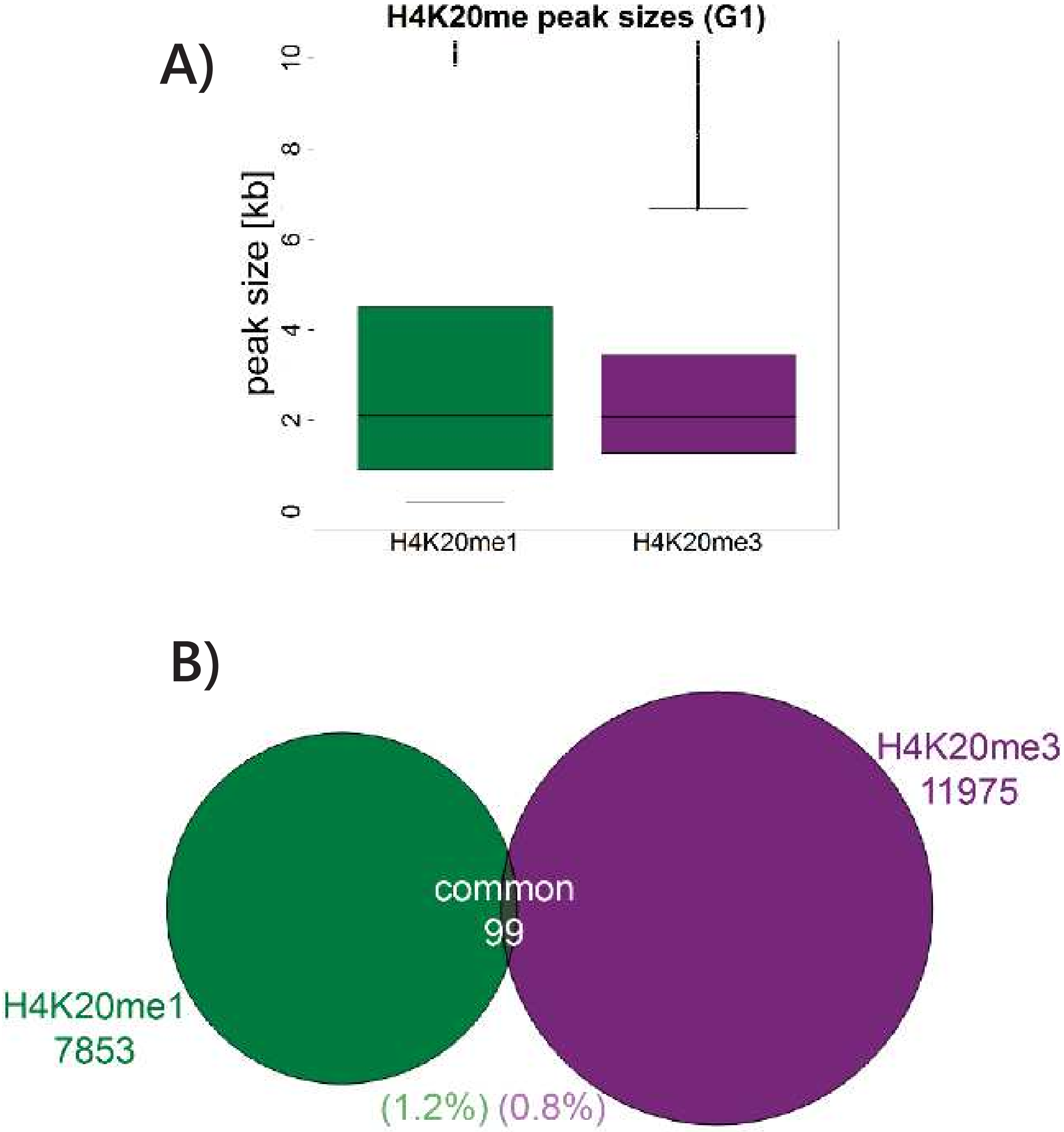
APPENDIX FIGURE 20: OVERLAPS OF HOMER- (LEFT PANEL) OR T-PIC- (RIGHT PANEL) DEFINED PRE-RC WITH H3K4ME3 PEAKS IN G1 VS. S/G2. Left: HOMER-defined complexes, right: T-PIC-defined complexes. Venn diagram of overlap between G1- and S/G2-defined pre-RCs with H3K4me3 peaks. Overall counts are indicated. The percentage of defined pre-RC overlapping with H3K4me3 are specified in brackets.



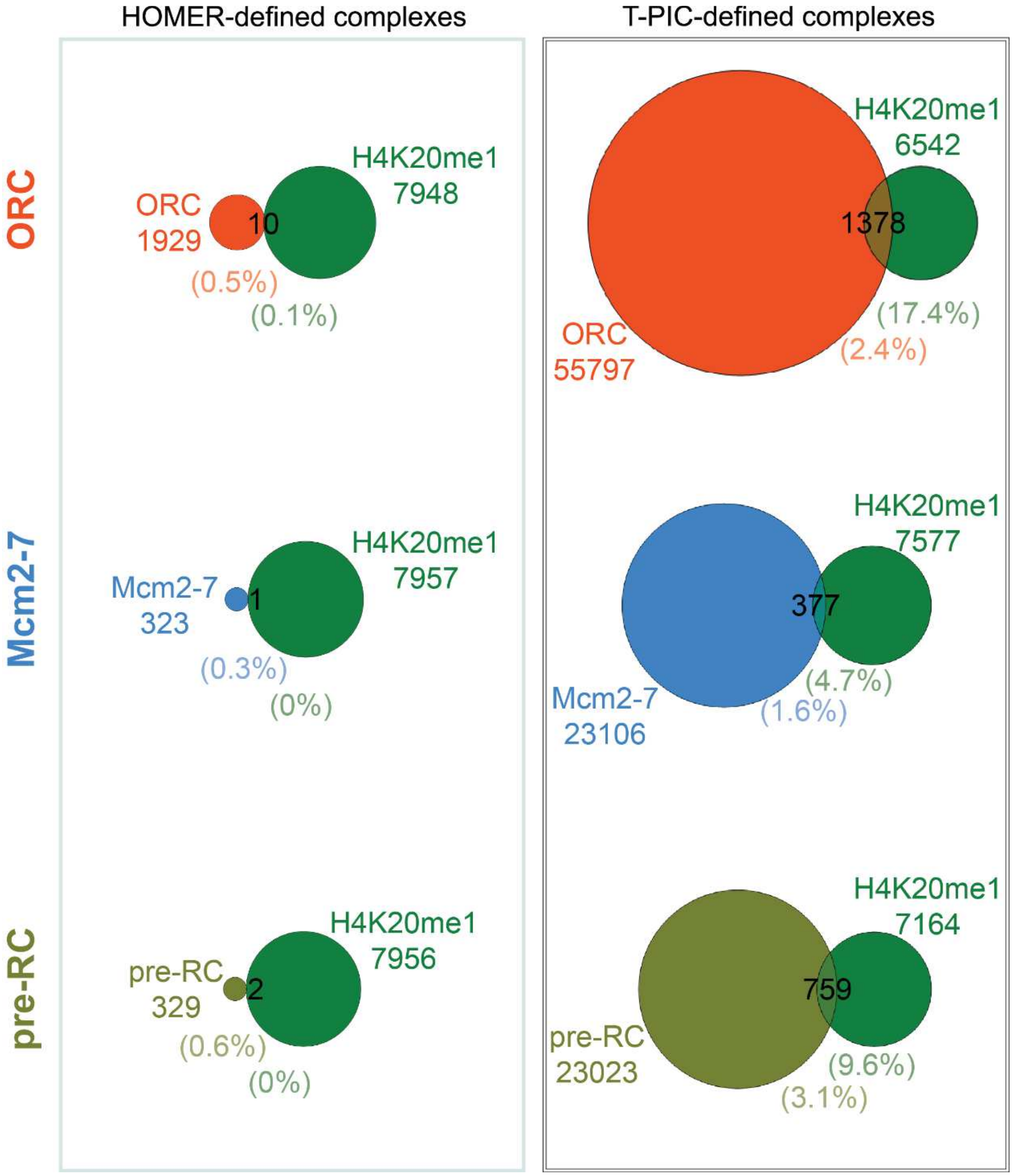
APPENDIX FIGURE 21: OVERLAPS OF HOMER- (LEFT PANEL) OR T-PIC- (RIGHT PANEL) DEFINED ORC AND MCM2-7 WITH H3K4ME3 PEAKS IN G1 VS. S/G2. Left: HOMER-defined complexes, right: T-PIC-defined complexes. Venn diagram of overlap between G1- and S/G2-defined complexes with H3K4me3 peaks. Overall counts are indicated. The percentage of defined complexes overlapping with H3K4me3 is specified in brackets.



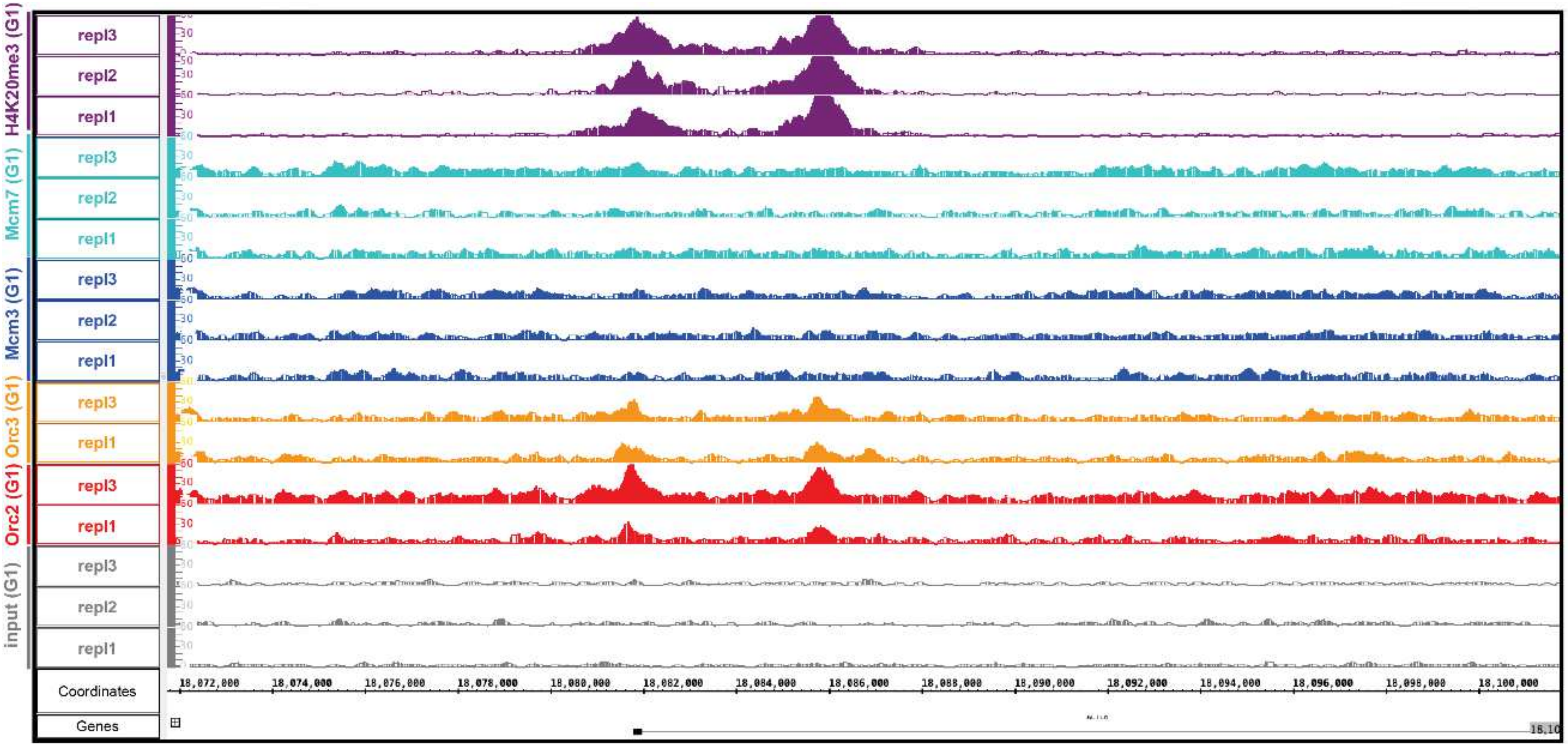
APPENDIX FIGURE 22: GENOMIC PRE-RC DISTRIBUTION IN G1 VS. S/G2. A) HOMER-DEFINED PRE-RCS. B) T-PIC-DEFINED PRE-RCS. The distribution in G1 is plotted in rainbow color. The S/G2 distribution is overlaid by gray bars.



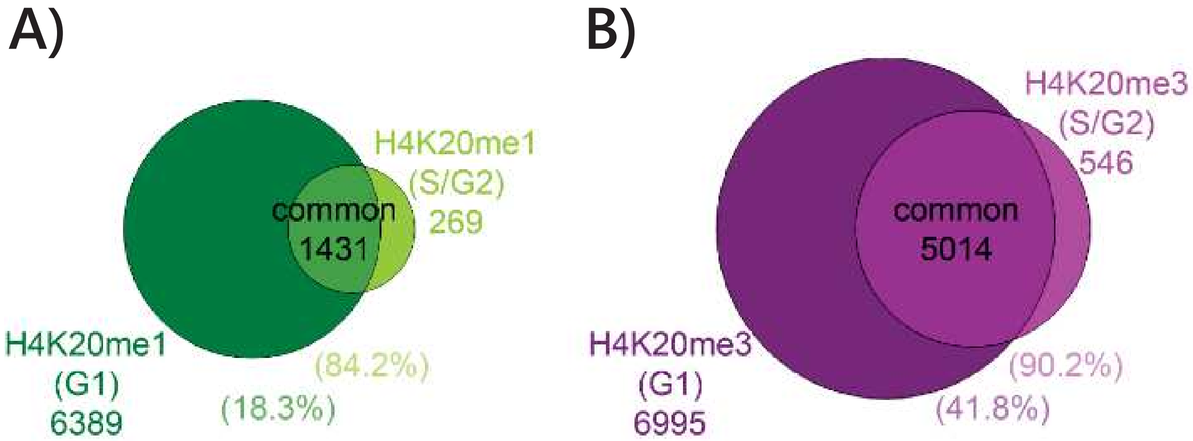
APPENDIX FIGURE 23: H4K20ME1 AND –ME3 POSITIONS ARE MUTUALLY EXCLUSIVE. A) H4K20ME1 AND –ME3 PEAK SIZES REPRESENTED IN A BOXLOT. Thick line shows the median, the box is the distribution from the first to the third quartile, the whiskers indicate the smallest and largest value without being an outlier. Outliers represented by dots. B) OVERLAP OF H4K20ME1 AND –ME3 PEAK POSITIONS. Venn diagram of overlapping H4K20me1 and H4K20me3 peaks. Overall counts are indicated. The percentage of overlapping proportions are specified below in brackets.



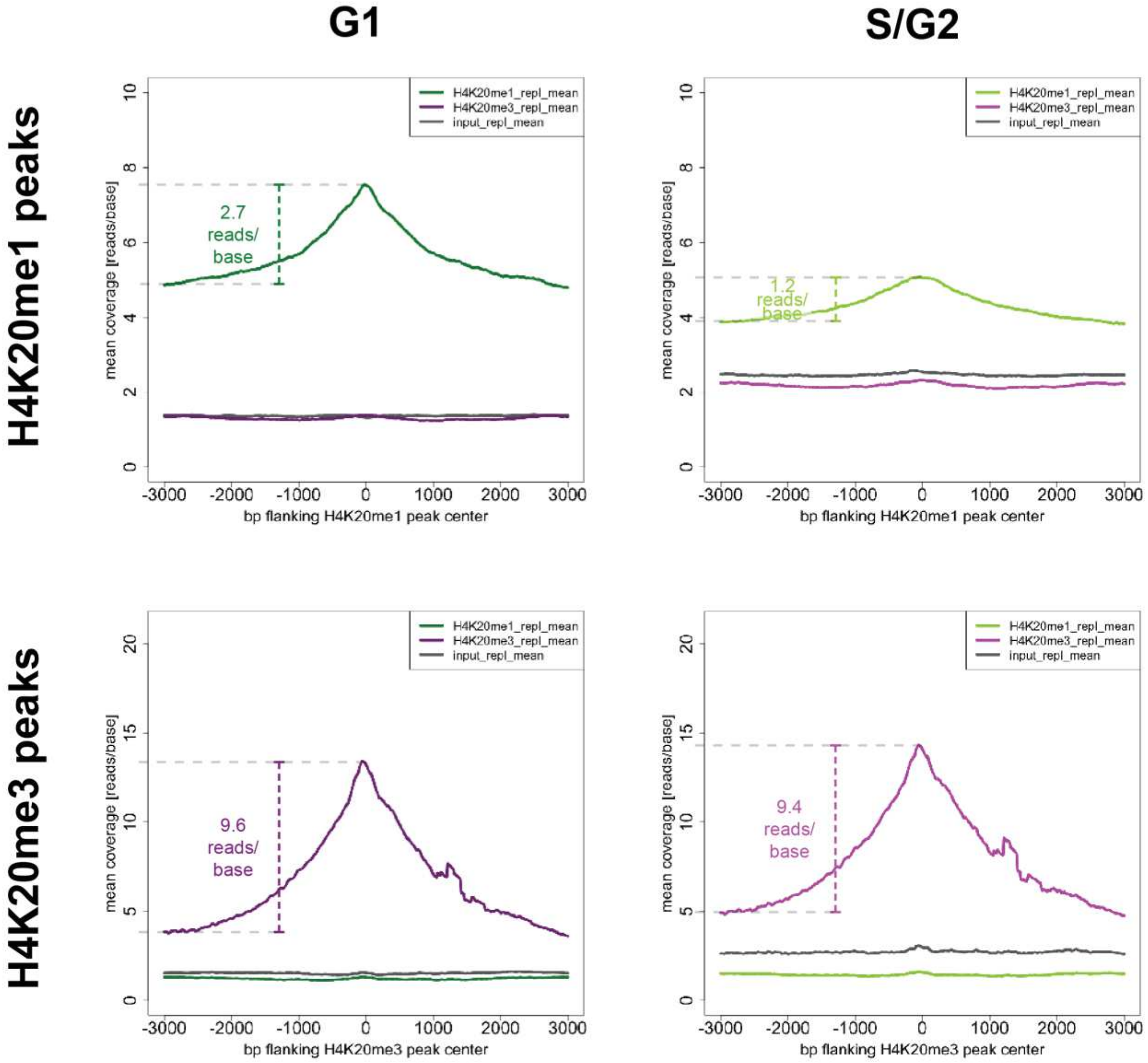
APPENDIX FIGURE 24: HOMER- (LEFT PANEL) AND T-PIC- (RIGHT PANEL) DEFINED COMPLEXES HARDLY OVERLAP WITH H4K20ME1. Left: HOMER-defined complexes, right: T-PIC-defined complexes. Venn diagram of overlap between HOMER- and T-PIC-defined ORC, Mcm2-7 and pre-RC with H4K20me1 peaks. Overall counts are indicated. The percentages of overlapping proportions are specified below in brackets.



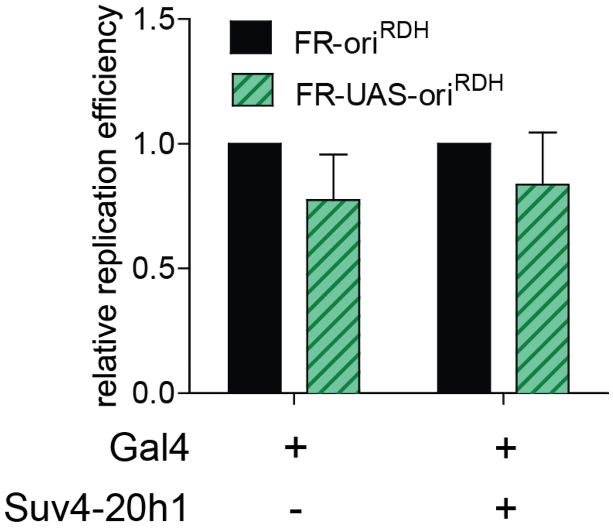
APPENDIX FIGURE 25: IGB VISUALIZATION OF ORC AND H4K20ME3 COLOCALIZATION. Input: grey, Orc2: red, Orc3: orange, Mcm3: blue, Mcm7: turquois, H4K20me3: violet; [chr1: 18071736-18101785].



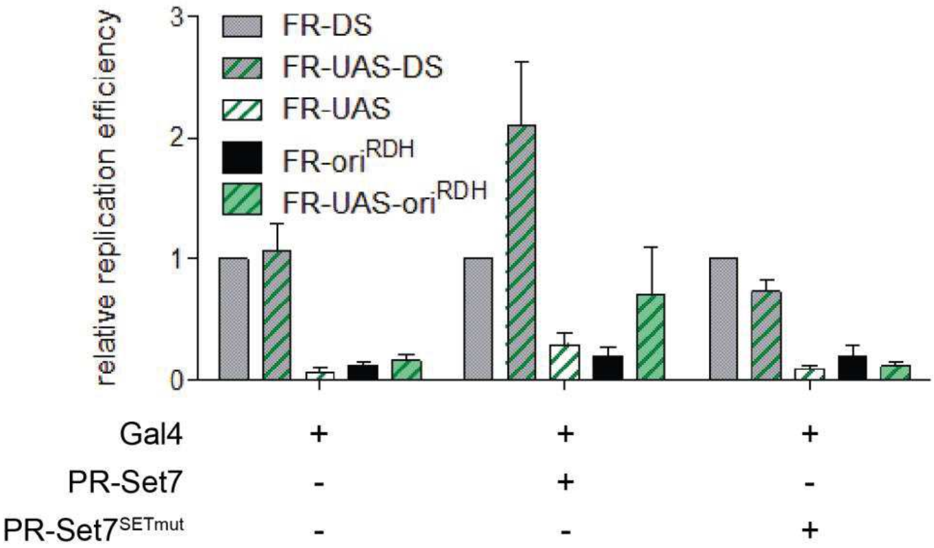
APPENDIX FIGURE 26: COMPARISON OF H4K20ME1 AND –ME3 POSITIONS IN G1 AND S/G2. A) H4K20ME1. B) H4K20ME3. Venn diagram of overlap between G1 and S/G2 determined H4K20me peaks. Overall counts are indicated. The percentage of overlapping proportions are specified below in brackets.



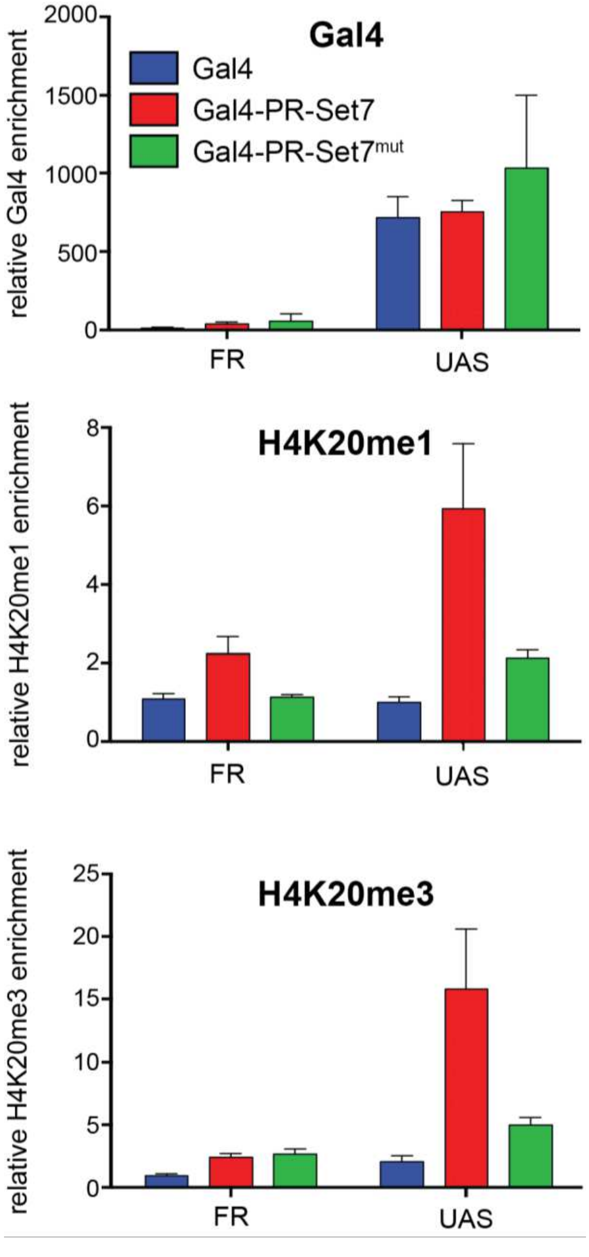
APPENDIX FIGURE 27: H4K20ME1 AND –ME3 COVERAGE AT H4K20ME1 AND –ME 3 PEAKS (G1 VS. S/G2). Coverage was calculated as number of reads/base within a 6 kb window around H4K20me1/-me3 peak center. H4K20me1 coverage in green, H4K20me3 coverage in violet, S/G2 coverages colored lighter.



APPENDIX FIGURE 28: SUV4-20H1 TARGETING HAS NO EFFECT ON PLASMID REPLICATION. QUANTIFICATION OF FR-ORI^{RDH} AND FR-UAS-ORI^{RDH} PLASMIDS. Plasmid abundance assays in HEK293 EBNA1^{+} Gal4/ Gal4-Suv4-20h1 cell lines as indicated: FR-ori^{RDH} replication efficiency was arbitrarily defined as 1. Data are means ± SEM (n=3).



APPENDIX FIGURE 29: DIRECT COMPARISON OF ALL REPORTER PLASMID REPLICATION EFFICIENCIES RELATIVE TO FR-DS. Plasmid abundance assays in HEK293 EBNA1$^+$ Gal4/ Gal4-PR-Set7/ Gal4-PR-Set7^{SETmut} cell lines as indicated: FR-DS replication efficiency was arbitrarily defined as 1. Data are means ± SEM (n=4).



APPENDIX FIGURE 30: PR-SET7 TARGETING TO FR-UAS LED TO H4K20ME1 INDUCTION AND CONVERSION INTO H4K20ME3. ChIP-qPCR analysis at FR and UAS sequences of FR-UAS reporter plasmids transfected in the indicated cell lines. Fold enrichment relative to IgG. Data are means ± SEM (n=3).

BIBLIOGRAPHY

Abbas, Tarek, Mignon A. Keaton, and Anindya Dutta. 2013. "Genomic Instability in Cancer." *Cold Spring Harbor Perspectives in Biology* 5 (3).

Abbas, Tarek, Etsuko Shibata, Jonghoon Park, Sudhakar Jha, Neerja Karnani, and Anindya Dutta. 2010. "CRL4(Cdt2) Regulates Cell Proliferation and Histone Gene Expression by Targeting PR-Set7/Set8 for Degradation." *Molecular Cell* 40 (1): 9–21.

Abdurashidova, G., M. Deganuto, R. Klima, S. Riva, G. Biamonti, M. Giacca, and A. Falaschi. 2000. "Start Sites of Bidirectional DNA Synthesis at the Human Lamin B2 Origin." *Science (New York, N.Y.)* 287 (5460): 2023–26.

Ballabeni, Andrea, Marina Melixetian, Raffaella Zamponi, Laura Masiero, Federica Marinoni, and Kristian Helin. 2004. "Human Geminin Promotes Pre-RC Formation and DNA Replication by Stabilizing CDT1 in Mitosis." *The EMBO Journal* 23 (15): 3122–32.

Beck, David B., Adam Burton, Hisanobu Oda, Céline Ziegler-Birling, Maria-Elena Torres-Padilla, and Danny Reinberg. 2012. "The Role of PR-Set7 in Replication Licensing Depends on Suv4-20h." *Genes & Development* 26 (23): 2580–89.

Bell, Stephen P., and Jon M. Kaguni. 2013. "Helicase Loading at Chromosomal Origins of Replication." *Cold Spring Harbor Perspectives in Biology* 5 (6).

Besnard, Emilie, Amélie Babled, Laure Lapasset, Ollivier Milhavet, Hugues Parrinello, Christelle Dantec, Jean-Michel Marin, and Jean-Marc Lemaitre. 2012. "Unraveling Cell Type–specific and Reprogrammable Human Replication Origin Signatures Associated with G-Quadruplex Consensus Motifs." *Nature Structural & Molecular Biology* 19 (8): 837–44.

Bleichert, Franziska, Michael R. Botchan, and James M. Berger. 2015. "Crystal Structure of the Eukaryotic Origin Recognition Complex." *Nature* 519 (7543): 321–26.

Blow, J. Julian, and Anindya Dutta. 2005. "Preventing Re-Replication of Chromosomal DNA." *Nature Reviews. Molecular Cell Biology* 6 (6): 476–86.

Blow, J. Julian, Xin Quan Ge, and Dean A. Jackson. 2011. "How Dormant Origins Promote Complete Genome Replication." *Trends in Biochemical Sciences* 36 (8): 405–14.

Brustel, Julien, Mathieu Tardat, Olivier Kirsh, Charlotte Grimaud, and Eric Julien. 2011. "Coupling Mitosis to DNA Replication: The Emerging Role of the Histone H4-Lysine 20 Methyltransferase PR-Set7." *Trends in Cell Biology* 21 (8): 452–60.

Cadoret, Jean-Charles, Françoise Meisch, Vahideh Hassan-Zadeh, Isabelle Luyten, Claire Guillet, Laurent Duret, Hadi Quesneville, and Marie-Noëlle Prioleau. 2008. "Genome-Wide Studies Highlight Indirect Links between Human Replication Origins and Gene Regulation." *Proceedings of the National Academy of Sciences of the United States of America* 105 (41): 15837–42.

Cayrou, Christelle, Benoit Ballester, Isabelle Peiffer, Romain Fenouil, Philippe Coulombe, Jean-Christophe Andrau, Jacques van Helden, and Marcel Méchali. 2015. "The Chromatin Environment Shapes DNA Replication Origin Organization and Defines Origin Classes." *Genome Research* 25 (12): 1873–85.

Cayrou, Christelle, Philippe Coulombe, Alice Vigneron, Slavica Stanojcic, Olivier Ganier, Isabelle Peiffer, Eric Rivals, *et al.* 2011. "Genome-Scale Analysis of Metazoan Replication Origins Reveals Their Organization in Specific but Flexible Sites Defined by Conserved Features." *Genome Research* 21 (9): 1438–49.

Centore, Richard C., Courtney G. Havens, Amity L. Manning, Ju-Mei Li, Rachel Litman Flynn, Alice Tse, Jianping Jin, Nicholas J. Dyson, Johannes C. Walter, and Lee Zou. 2010. "CRL4(Cdt2)-Mediated Destruction of the Histone Methyltransferase Set8 Prevents Premature Chromatin Compaction in S Phase." *Molecular Cell* 40 (1): 22–33.

Chakraborty, Arindam, Zhen Shen, and Supriya G. Prasanth. 2011. "'ORCanization' on Heterochromatin: Linking DNA Replication Initiation to Chromatin Organization." *Epigenetics* 6 (6): 665–70.

Ciabrelli, Filippo, and Giacomo Cavalli. 2015. "Chromatin-Driven Behavior of Topologically Associating Domains." *Journal of Molecular Biology*, Functional Relevance and Dynamics of Nuclear Organization, 427 (3): 608–25.

Clijsters, Linda, and Rob Wolthuis. 2014. "PIP-Box-Mediated Degradation Prohibits Re-Accumulation of Cdc6 during S Phase." *Journal of Cell Science* 127 (Pt 6): 1336–45.

Das, Shankar P., Tyler Borrman, Victor W. T. Liu, Scott C.-H. Yang, John Bechhoefer, and Nicholas Rhind. 2015. "Replication Timing Is Regulated by the Number of MCMs Loaded at Origins." *Genome Research* 25 (12): 1886–92.

Das, Shankar P., and Nicholas Rhind. 2016. "How and Why Multiple MCMs Are Loaded at Origins of DNA Replication." *BioEssays* 38 (7): 613–17.

De Carli, Francesco, Vincent Gaggioli, Gaël A. Millot, and Olivier Hyrien. 2016. "Single-Molecule, Antibody-Free Fluorescent Visualisation of Replication Tracts along Barcoded DNA Molecules." *The International Journal of Developmental Biology*, May.

Dellino, Gaetano Ivan, Davide Cittaro, Rossana Piccioni, Lucilla Luzi, Stefania Banfi, Simona Segalla, Matteo Cesaroni, Ramiro Mendoza-Maldonado, Mauro Giacca, and Pier Giuseppe Pelicci. 2013. "Genome-Wide Mapping of Human DNA-Replication Origins: Levels of Transcription at ORC1 Sites Regulate Origin Selection and Replication Timing." *Genome Research* 23 (1): 1–11.

DePamphilis, Melvin L. 2005. "Cell Cycle Dependent Regulation of the Origin Recognition Complex." *Cell Cycle (Georgetown, Tex.)* 4 (1): 70–79.

DePamphilis, Melvin L., J. Julian Blow, Soma Ghosh, Tapas Saha, Kohji Noguchi, and Alex Vassilev. 2006. "Regulating the Licensing of DNA Replication Origins in Metazoa." *Current Opinion in Cell Biology* 18 (3): 231–39.

Depamphilis, Melvin L., Christelle M. de Renty, Zakir Ullah, and Chrissie Y. Lee. 2012. "'The Octet': Eight Protein Kinases That Control Mammalian DNA Replication." *Systems Biology* 3: 368.

Dillin, A., and J. Rine. 1997. "Separable Functions of ORC5 in Replication Initiation and Silencing in Saccharomyces Cerevisiae." *Genetics* 147 (3): 1053–62.

Ding, Queying, and David M. MacAlpine. 2011. "Defining the Replication Program through the Chromatin Landscape." *Critical Reviews in Biochemistry and Molecular Biology* 46 (2): 165–79.

Donovan, S., J. Harwood, L. S. Drury, and J. F. Diffley. 1997. "Cdc6p-Dependent Loading of Mcm Proteins onto Pre-Replicative Chromatin in Budding Yeast." *Proceedings of the National Academy of Sciences of the United States of America* 94 (11): 5611–16.

Fachinetti, Daniele, Rodrigo Bermejo, Andrea Cocito, Simone Minardi, Yuki Katou, Yutaka Kanoh, Katsuhiko Shirahige, Anna Azvolinsky, Virginia A. Zakian, and Marco Foiani. 2010. "Replication Termination at Eukaryotic Chromosomes Is Mediated by Top2 and Occurs at Genomic Loci Containing Pausing Elements." *Molecular Cell* 39 (4): 595–605..

Feng, Jianxing, Tao Liu, and Yong Zhang. 2011. "Using MACS to Identify Peaks from ChIP-Seq Data." *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et al.]* CHAPTER (June): Unit2.14.

Fenouil, Romain, Pierre Cauchy, Frederic Koch, Nicolas Descostes, Joaquin Zacarias Cabeza, Charlène Innocenti, Pierre Ferrier, *et al.* 2012. "CpG Islands and GC Content Dictate Nucleosome Depletion in a Transcription-Independent Manner at Mammalian Promoters." *Genome Research* 22 (12): 2399–2408.

Field, Yair, Noam Kaplan, Yvonne Fondufe-Mittendorf, Irene K. Moore, Eilon Sharon, Yaniv Lubling, Jonathan Widom, and Eran Segal. 2008. "Distinct Modes of Regulation by Chromatin Encoded through Nucleosome Positioning Signals." *PLoS Computational Biology* 4 (11): e1000216.

Fragkos, Michalis, Olivier Ganier, Philippe Coulombe, and Marcel Méchali. 2015. "DNA Replication Origin Activation in Space and Time." *Nature Reviews Molecular Cell Biology* 16 (6): 360–74.

Francis, Laura I., John C.W. Randell, Thomas J. Takara, Lilen Uchima, and Stephen P. Bell. 2009. "Incorporation into the Prereplicative Complex Activates the Mcm2–7 Helicase for Cdc7–Dbf4 Phosphorylation." *Genes & Development* 23 (5): 643–54.

Fu, Haiqing, Emilie Besnard, Romain Desprat, Michael Ryan, Malik Kahli, Jean-Marc Lemaitre, and Mirit I. Aladjem. 2014. "Mapping Replication Origin Sequences in Eukaryotic Chromosomes." *Current Protocols in Cell Biology / Editorial Board, Juan S. Bonifacino ... [et al.]* 65: 22.20.1-17.

Fu, Haiqing, Alika K. Maunakea, Melvenia M. Martin, Liang Huang, Ya Zhang, Michael Ryan, RyangGuk Kim, Chii Meil Lin, Keji Zhao, and Mirit I. Aladjem. 2013. "Methylation of Histone H3 on Lysine 79 Associates with a Group of Replication Origins and Helps Limit DNA Replication Once per Cell Cycle." Edited by Christopher E. Pearson. *PLoS Genetics* 9 (6): e1003542.

Gambus, Agnieszka, Richard C. Jones, Alberto Sanchez-Diaz, Masato Kanemaki, Frederick van Deursen, Ricky D. Edmondson, and Karim Labib. 2006. "GINS Maintains Association of Cdc45 with MCM in Replisome Progression Complexes at Eukaryotic DNA Replication Forks." *Nature Cell Biology* 8 (4): 358–66.

Geng, Yan, Young-Mi Lee, Markus Welcker, Jherek Swanger, Agnieszka Zagozdzon, Joel D. Winer, James M. Roberts, Philipp Kaldis, Bruce E. Clurman, and Piotr

Sicinski. 2007. “Kinase-Independent Function of Cyclin E.” *Molecular Cell* 25 (1): 127–39.

Gerhardt, Jeannine, Samira Jafar, Mark-Peter Spindler, Elisabeth Ott, and Aloys Schepers. 2006. “Identification of New Human Origins of DNA Replication by an Origin-Trapping Assay.” *Molecular and Cellular Biology* 26 (20): 7731–46.

Ghosh, Maloy, Michael Kemp, Guoqi Liu, Marion Ritzi, Aloys Schepers, and Michael Leffak. 2006. “Differential Binding of Replication Proteins across the Human c-Myc Replicator.” *Molecular and Cellular Biology* 26 (14): 5270–83.

Giri, Sumanprava, Vasudha Aggarwal, Julien Pontis, Zhen Shen, Arindam Chakraborty, Abid Khan, Craig Mizzen, *et al.* 2015. “The preRC Protein ORCA Organizes Heterochromatin by Assembling Histone H3 Lysine 9 Methyltransferases on Chromatin.” *eLife* 4.

Giri, Sumanprava, and Supriya G. Prasanth. 2015. “Association of ORCA/LRWD1 with Repressive Histone Methyl Transferases Mediates Heterochromatin Organization.” *Nucleus (Austin, Tex.)* 6 (6): 435–41.

Gonzalez-Sandoval, Adriana, and Susan M. Gasser. 2016. “On TADs and LADs: Spatial Control Over Gene Expression.” *Trends in Genetics* 32 (8): 485–95.

Gros, Julien, Charanya Kumar, Gerard Lynch, Tejas Yadav, Iestyn Whitehouse, and Dirk Remus. 2015. “Post-Licensing Specification of Eukaryotic Replication Origins by Facilitated Mcm2-7 Sliding along DNA.” *Molecular Cell* 60 (5): 797–807.

Guelen, Lars, Ludo Pagie, Emilie Brasset, Wouter Meuleman, Marius B. Faza, Wendy Talhout, Bert H. Eussen, *et al.* 2008. “Domain Organization of Human Chromosomes Revealed by Mapping of Nuclear Lamina Interactions.” *Nature* 453 (7197): 948–51.

Hammerschmidt, Wolfgang, and Bill Sugden. 2013. “Replication of Epstein-Barr Viral DNA.” *Cold Spring Harbor Perspectives in Biology* 5 (1): a013029.

Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. “Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities.” *Molecular Cell* 38 (4): 576–89.

Herrera, M. Carmen, Silvia Tognetti, Alberto Riera, Juergen Zech, Pippa Clarke, Alejandra Fernández-Cid, and Christian Speck. 2015. “A Reconstituted System Reveals How Activating and Inhibitory Interactions Control DDK Dependent Assembly of the Eukaryotic Replicative Helicase.” *Nucleic Acids Research* 43 (21): 10238–50.

Hirt, B. 1966. “Evidence for Semiconservative Replication of Circular Polyoma DNA.” *Proceedings of the National Academy of Sciences of the United States of America* 55 (4): 997–1004.

Hower, Valerie, Steven N. Evans, and Lior Pachter. 2011. “Shape-Based Peak Identification for ChIP-Seq.” *BMC Bioinformatics* 12: 15.

Hua, X. H., and J. Newport. 1998. “Identification of a Preinitiation Step in DNA Replication That Is Independent of Origin Recognition Complex and cdc6, but Dependent on cdk2.” *The Journal of Cell Biology* 140 (2): 271–81.

Hyrien, Olivier. 2016. “How MCM Loading and Spreading Specify Eukaryotic DNA Replication Initiation Sites.” *F1000Research* 5.

Hyrien, Olivier, Kathrin Marheineke, and Arach Goldar. 2003. “Paradoxes of Eukaryotic DNA Replication: MCM Proteins and the Random Completion Problem.” *BioEssays* 25 (2): 116–25.

Iizuka, Masayoshi, Yoshihisa Takahashi, Craig A. Mizzen, Richard G. Cook, Masatoshi Fujita, C. David Allis, Henry F. Frierson, Toshio Fukusato, and M. Mitchell Smith. 2009. "Histone Acetyltransferase Hbo1: Catalytic Activity, Cellular Abundance, and Links to Primary Cancers." *Gene* 436 (1–2): 108–14.

Ilves, Ivar, Tatjana Petojevic, James J. Pesavento, and Michael R. Botchan. 2010. "Activation of the MCM2-7 Helicase by Association with Cdc45 and GINS Proteins." *Molecular Cell* 37 (2): 247–58.

Kimura, Hiroshi. 2013. "Histone Modifications for Human Epigenome Analysis." *Journal of Human Genetics* 58 (7): 439–45.

Kouzarides, Tony. 2007. "Chromatin Modifications and Their Function." *Cell* 128 (4): 693–705.

Kuipers, Marjorie A., Timothy J. Stasevich, Takayo Sasaki, Korey A. Wilson, Kristin L. Hazelwood, James G. McNally, Michael W. Davidson, and David M. Gilbert. 2011. "Highly Stable Loading of Mcm Proteins onto Chromatin in Living Cells Requires Replication to Unload." *The Journal of Cell Biology* 192 (1): 29–41.

Kuo, Alex J., Jikui Song, Peggie Cheung, Satoko Ishibe-Murakami, Sayumi Yamazoe, James K. Chen, Dinshaw J. Patel, and Or Gozani. 2012. "The BAH Domain of ORC1 Links H4K20me2 to DNA Replication Licensing and Meier-Gorlin Syndrome." *Nature* 484 (7392): 115–19.

Kylie, Kathleen, Julia Romero, Indeewari K. S. Lindamulage, James Knockleby, and Hoyun Lee. 2016. "Dynamic Regulation of Histone H3K9 Is Linked to the Switch between Replication and Transcription at the Dbf4 Origin-Promoter Locus." *Cell Cycle (Georgetown, Tex.)* 15 (17): 2321–35.

Langley, Alexander R., Stefan Gräf, James C. Smith, and Torsten Krude. 2016. "Genome-Wide Identification and Characterisation of Human DNA Replication Origins by Initiation Site Sequencing (Ini-Seq)." *Nucleic Acids Research*, September.

Leatherwood, Janet, and Amit Vas. 2003. "Connecting ORC and Heterochromatin: Why?" *Cell Cycle (Georgetown, Tex.)* 2 (6): 573–75.

Lee, Kyung Yong, Sung Woong Bang, Sang Wook Yoon, Seung-Hoon Lee, Jong-Bok Yoon, and Deog Su Hwang. 2012. "Phosphorylation of ORC2 Protein Dissociates Origin Recognition Complex from Chromatin and Replication Origins." *The Journal of Biological Chemistry* 287 (15): 11891–98.

Lengronne, Armelle, and Philippe Pasero. 2014. "Closing the MCM Cycle at Replication Termination Sites." *EMBO Reports* 15 (12): 1226–27.

Lombraña, Rodrigo, Alba Álvarez, José Miguel Fernández-Justel, Ricardo Almeida, César Poza-Carrión, Fábia Gomes, Arturo Calzada, José María Requena, and María Gómez. 2016. "Transcriptionally Driven DNA Replication Program of the Human Parasite Leishmania Major." *Cell Reports* 16 (6): 1774–86.

Lõoke, Marko, Jüri Reimand, Tiina Sedman, Juhan Sedman, Lari Järvinen, Signe Värv, Kadri Peil, Kersti Kristjuhan, Jaak Vilo, and Arnold Kristjuhan. 2010. "Relicensing of Transcriptionally Inactivated Replication Origins in Budding Yeast." *The Journal of Biological Chemistry* 285 (51): 40004–11.

Lucas, I., M. Chevrier-Miller, J. M. Sogo, and O. Hyrien. 2000. "Mechanisms Ensuring Rapid and Complete DNA Replication despite Random Initiation in Xenopus Early Embryos." *Journal of Molecular Biology* 296 (3): 769–86.

Ludwig, Tenneille E., Veit Bergendahl, Mark E. Levenstein, Junying Yu, Mitchell D. Probasco, and James A. Thomson. 2006. "Feeder-Independent Culture of Human Embryonic Stem Cells." *Nature Methods* 3 (8): 637–46.

Marechal, V., A. Dehee, R. Chikhi-Brachet, T. Piolot, M. Coppey-Moisan, and J. C. Nicolas. 1999. "Mapping EBNA-1 Domains Involved in Binding to Metaphase Chromosomes." *Journal of Virology* 73 (5): 4385–92.

Maric, Marija, Timurs Maculins, Giacomo De Piccoli, and Karim Labib. 2014. "Cdc48 and a Ubiquitin Ligase Drive Disassembly of the CMG Helicase at the End of DNA Replication." *Science (New York, N.Y.)* 346 (6208): 1253596.

Martin, Melvenia M., Michael Ryan, RyangGuk Kim, Anna L. Zakas, Haiqing Fu, Chii Mei Lin, William C. Reinhold, *et al.* 2011. "Genome-Wide Depletion of Replication Initiation Events in Highly Transcribed Regions." *Genome Research* 21 (11): 1822–32.

Méchali, Marcel. 2010. "Eukaryotic DNA Replication Origins: Many Choices for Appropriate Answers." *Nature Reviews Molecular Cell Biology* 11 (10): 728–38.

Mesner, Larry D., Emily L. Crawford, and Joyce L. Hamlin. 2006. "Isolating Apparently Pure Libraries of Replication Origins from Complex Genomes." *Molecular Cell* 21 (5): 719–26.

Mesner, Larry D., Veena Valsakumar, Marcin Cieślik, Rebecca Pickin, Joyce L. Hamlin, and Stefan Bekiranov. 2013. "Bubble-Seq Analysis of the Human Genome Reveals Distinct Chromatin-Mediated Mechanisms for Regulating Early- and Late-Firing Origins." *Genome Research* 23 (11): 1774–88.

Micklem, G., A. Rowley, J. Harwood, K. Nasmyth, and J. F. Diffley. 1993. "Yeast Origin Recognition Complex Is Involved in DNA Replication and Transcriptional Silencing." *Nature* 366 (6450): 87–89.

Miotto, Benoit, Zhe Ji, and Kevin Struhl. 2016. "Selectivity of ORC Binding Sites and the Relation to Replication Timing, Fragile Sites, and Deletions in Cancers." *Proceedings of the National Academy of Sciences of the United States of America* 113 (33): E4810-4819.

Miotto, Benoit, and Kevin Struhl. 2010. "HBO1 Histone Acetylase Activity Is Essential for DNA Replication Licensing and Inhibited by Geminin." *Molecular Cell* 37 (1): 57–66.

Moreno, Sara Priego, Rachael Bailey, Nicholas Campion, Suzanne Herron, and Agnieszka Gambus. 2014. "Polyubiquitylation Drives Replisome Disassembly at the Termination of DNA Replication." *Science* 346 (6208): 477–81.

Mori, Saori, and Katsuhiko Shirahige. 2007. "Perturbation of the Activity of Replication Origin by Meiosis-Specific Transcription." *The Journal of Biological Chemistry* 282 (7): 4447–52.

Mott, Melissa L., and James M. Berger. 2007. "DNA Replication Initiation: Mechanisms and Regulation in Bacteria." *Nature Reviews. Microbiology* 5 (5): 343–54.

Musiałek, Marcelina W, and Dorota Rybaczek. 2015. "Behavior of Replication Origins in Eukaryota – Spatio-Temporal Dynamics of Licensing and Firing." *Cell Cycle* 14 (14): 2251–64.

Nieduszynski, Conrad A., J. Julian Blow, and Anne D. Donaldson. 2005. "The Requirement of Yeast Replication Origins for Pre-Replication Complex Proteins Is Modulated by Transcription." *Nucleic Acids Research* 33 (8): 2410–20.

Oda, Hisanobu, Michael R. Hübner, David B. Beck, Michiel Vermeulen, Jerard Hurwitz, David L. Spector, and Danny Reinberg. 2010. "Regulation of the Histone H4 Monomethylase PR-Set7 by CRL4(Cdt2)-Mediated PCNA-Dependent Degradation during DNA Damage." *Molecular Cell* 40 (3): 364–76.

Pak, D. T., M. Pflumm, I. Chesnokov, D. W. Huang, R. Kellum, J. Marr, P. Romanowski, and M. R. Botchan. 1997. "Association of the Origin Recognition Complex with Heterochromatin and HP1 in Higher Eukaryotes." *Cell* 91 (3): 311–23.

Papior, Peer. 2010. "Die Ausbildung von Pre-Replikationskomplexen Im Epstein-Barr-Virus Und Dem Menschen." Text.PhDThesis, Ludwig-Maximilians-Universität München. https://edoc.ub.uni-muenchen.de/13181/.

Papior, Peer, José M. Arteaga-Salas, Thomas Günther, Adam Grundhoff, and Aloys Schepers. 2012. "Open Chromatin Structures Regulate the Efficiencies of Pre-RC Formation and Replication Initiation in Epstein-Barr Virus." *The Journal of Cell Biology* 198 (4): 509–28.

Pesavento, James J., Hongbo Yang, Neil L. Kelleher, and Craig A. Mizzen. 2008. "Certain and Progressive Methylation of Histone H4 at Lysine 20 during the Cell Cycle." *Molecular and Cellular Biology* 28 (1): 468–86.

Petryk, Nataliya, Malik Kahli, Yves d'Aubenton-Carafa, Yan Jaszczyszyn, Yimin Shen, Maud Silvain, Claude Thermes, Chun-Long Chen, and Olivier Hyrien. 2016. "Replication Landscape of the Human Genome." *Nature Communications* 7 (January).

Picard, Franck, Jean-Charles Cadoret, Benjamin Audit, Alain Arneodo, Adriana Alberti, Christophe Battail, Laurent Duret, and Marie-Noelle Prioleau. 2014. "The Spatiotemporal Program of DNA Replication Is Associated with Specific Combinations of Chromatin Marks in Human Cells." *PLoS Genetics* 10 (5).

Pines, J. 1999. "Four-Dimensional Control of the Cell Cycle." *Nature Cell Biology* 1 (3): E73–79.

Pope, Benjamin D., and David M. Gilbert. 2013. "The Replication Domain Model: Regulating Replicon Firing in the Context of Large-Scale Chromosome Architecture." *Journal of Molecular Biology* 425 (23).

Pope, Benjamin D., Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas, Daniel L. Vera, *et al.* 2014. "Topologically-Associating Domains Are Stable Units of Replication-Timing Regulation." *Nature* 515 (7527): 402–5.

Powell, Sara K., Heather K. MacAlpine, Joseph A. Prinz, Yulong Li, Jason A. Belsky, and David M. MacAlpine. 2015. "Dynamic Loading and Redistribution of the Mcm2-7 Helicase Complex through the Cell Cycle." *The EMBO Journal* 34 (4): 531.

Remus, Dirk, Fabienne Beuron, Gökhan Tolun, Jack D. Griffith, Edward P. Morris, and John F.X. Diffley. 2009. "Concerted Loading of Mcm2-7 Double Hexamers Around DNA during DNA Replication Origin Licensing." *Cell* 139 (4): 719–30.

Ritzi, Marion, Kristina Tillack, Jeannine Gerhardt, Elisabeth Ott, Sibille Humme, Elisabeth Kremmer, Wolfgang Hammerschmidt, and Aloys Schepers. 2003. "Complex Protein-DNA Dynamics at the Latent Origin of DNA Replication of Epstein-Barr Virus." *Journal of Cell Science* 116 (19): 3971–84.

Rivera-Mulia, Juan Carlos, and David M. Gilbert. 2016a. "Replication Timing and Transcriptional Control: Beyond Cause and Effect-Part III." *Current Opinion in Cell Biology* 40 (June): 168–78.

———. 2016b. "Replicating Large Genomes: Divide and Conquer." *Molecular Cell* 62 (5): 756–65.

Saredi, Giulia, Hongda Huang, Colin M. Hammond, Constance Alabert, Simon Bekker-Jensen, Ignasi Forne, Nazaret Reverón-Gómez, *et al.* 2016. "H4K20me0 Marks Post-Replicative Chromatin and Recruits the TONSL–MMS22L DNA Repair Complex." *Nature* 534 (7609): 714–18.

Schaarschmidt, Daniel, Eva-Maria Ladenburger, Christian Keller, and Rolf Knippers. 2002. "Human Mcm Proteins at a Replication Origin during the G1 to S Phase Transition." *Nucleic Acids Research* 30 (19): 4176–85.

Schepers, Aloys, and Peer Papior. 2010. "Why Are We Where We Are? Understanding Replication Origins and Initiation Sites in Eukaryotes Using ChIP-Approaches." *Chromosome Research* 18 (1): 63–77.

Schepers, Aloys, Marion Ritzi, Kristine Bousset, Elisabeth Kremmer, John L. Yates, Janet Harwood, John F.X. Diffley, and Wolfgang Hammerschmidt. 2001. "Human Origin Recognition Complex Binds to the Region of the Latent Origin of DNA Replication of Epstein–Barr Virus." *The EMBO Journal* 20 (16): 4588–4602.

Schotta, Gunnar, Monika Lachner, Kavitha Sarma, Anja Ebert, Roopsha Sengupta, Gunter Reuter, Danny Reinberg, and Thomas Jenuwein. 2004. "A Silencing Pathway to Induce H3-K9 and H4-K20 Trimethylation at Constitutive Heterochromatin." *Genes & Development* 18 (11): 1251–62.

Schotta, Gunnar, Roopsha Sengupta, Stefan Kubicek, Stephen Malin, Monika Kauer, Elsa Callén, Arkady Celeste, *et al.* 2008. "A Chromatin-Wide Transition to H4K20 Monomethylation Impairs Genome Integrity and Programmed DNA Rearrangements in the Mouse." *Genes & Development* 22 (15): 2048–61.

Sears, John, John Kolman, Geoffrey M. Wahl, and Ashok Aiyar. 2003. "Metaphase Chromosome Tethering Is Necessary for the DNA Synthesis and Maintenance of oriP Plasmids but Is Insufficient for Transcription Activation by Epstein-Barr Nuclear Antigen 1." *Journal of Virology* 77 (21): 11767–80.

Sears, John, Maki Ujihara, Samantha Wong, Christopher Ott, Jaap Middeldorp, and Ashok Aiyar. 2004. "The Amino Terminus of Epstein-Barr Virus (EBV) Nuclear Antigen 1 Contains AT Hooks That Facilitate the Replication and Partitioning of Latent EBV Genomes by Tethering Them to Cellular Chromosomes." *Journal of Virology* 78 (21): 11487–505.

Sequeira-Mendes, Joana, Ramón Díaz-Uriarte, Anwyn Apedaile, Derek Huntley, Neil Brockdorff, and María Gómez. 2009. "Transcription Initiation Activity Sets Replication Origin Efficiency in Mammalian Cells." *PLoS Genetics* 5 (4): e1000446.

Shareef, Mohammed M., RamaKrishna Badugu, and Rebecca Kellum. 2003. "HP1/ORC Complex and Heterochromatin Assembly." *Genetica* 117 (2–3): 127–34.

Shen, Zhen, Arindam Chakraborty, Ankur Jain, Sumanprava Giri, Taekjip Ha, Kannanganattu V. Prasanth, and Supriya G. Prasanth. 2012. "Dynamic Association of ORCA with Prereplicative Complex Components Regulates DNA Replication Initiation." *Molecular and Cellular Biology* 32 (15): 3107–20.

Shen, Zhen, Kizhakke M. Sathyan, Yijie Geng, Ruiping Zheng, Arindam Chakraborty, Brian Freeman, Fei Wang, Kannanganattu V. Prasanth, and Supriya G. Prasanth. 2010. "A WD-Repeat Protein Stabilizes ORC Binding to Chromatin." *Molecular Cell* 40 (1): 99–111.

Shin, Hyunjin, Tao Liu, Arjun K. Manrai, and X. Shirley Liu. 2009. "CEAS: Cis-Regulatory Element Annotation System." *Bioinformatics* 25 (19): 2605–6.

Smith, Owen K., RyanGuk Kim, Haiqing Fu, Melvenia M. Martin, Chii Mei Lin, Koichi Utani, Ya Zhang, *et al.* 2016. "Distinct Epigenetic Features of Differentiation-Regulated Replication Origins." *Epigenetics & Chromatin* 9: 18.

Sonneville, Remi, Matthieu Querenet, Ashley Craig, Anton Gartner, and J. Julian Blow. 2012. "The Dynamics of Replication Licensing in Live Caenorhabditis Elegans Embryos." *The Journal of Cell Biology* 196 (2): 233–46.

Soshnev, Alexey A., Steven Z. Josefowicz, and C. David Allis. 2016. "Greater Than the Sum of Parts: Complexity of the Dynamic Epigenome." *Molecular Cell* 62 (5): 681–94.

Tardat, Mathieu, Julien Brustel, Olivier Kirsh, Christine Lefevbre, Mary Callanan, Claude Sardet, and Eric Julien. 2010. "The Histone H4 Lys 20 Methyltransferase PR-Set7 Regulates Replication Origins in Mammalian Cells." *Nature Cell Biology* 12 (11): 1086–93.

Urban, John M., Michael S. Foulk, Cinzia Casella, and Susan A. Gerbi. 2015. "The Hunt for Origins of DNA Replication in Multicellular Eukaryotes." *F1000Prime Reports* 7 (March).

Valenzuela, Manuel S., Yidong Chen, Sean Davis, Fan Yang, Robert L. Walker, Sven Bilke, John Lueders, *et al.* 2011. "Preferential Localization of Human Origins of DNA Replication at the 5'-ends of Expressed Genes and at Evolutionarily Conserved DNA Sequences." *PloS One* 6 (5): e17308.

Valton, Anne-Laure, Vahideh Hassan-Zadeh, Ingrid Lema, Nicole Boggetto, Patrizia Alberti, Carole Saintomé, Jean-Francois Riou, and Marie-Noëlle Prioleau. 2014. "G4 Motifs Affect Origin Positioning and Efficiency in Two Vertebrate Replicators." *The EMBO Journal* 33 (7): 732–46.

Valton, Anne-Laure, and Marie-Noëlle Prioleau. 2016. "G-Quadruplexes in DNA Replication: A Problem or a Necessity?" *Trends in Genetics*. Accessed October 3.

Vashee, Sanjay, Christin Cvetic, Wenyan Lu, Pamela Simancek, Thomas J. Kelly, and Johannes C. Walter. 2003. "Sequence-Independent DNA Binding and Replication Initiation by the Human Origin Recognition Complex." *Genes & Development* 17 (15): 1894–1908.

Vermeulen, Michiel, H. Christian Eberl, Filomena Matarese, Hendrik Marks, Sergei Denissov, Falk Butter, Kenneth K. Lee, *et al.* 2010. "Quantitative Interaction Proteomics and Genome-Wide Profiling of Epigenetic Histone Marks and Their Readers." *Cell* 142 (6): 967–80.

Wilbanks, Elizabeth G., and Marc T. Facciotti. 2010. "Evaluation of Algorithm Performance in ChIP-Seq Peak Detection." *PLoS ONE* 5 (7).

Wilson, Korey A., Andrew G. Elefanty, Edouard G. Stanley, and David M. Gilbert. 2016. "Spatio-Temporal Re-Organization of Replication Foci Accompanies Replication Domain Consolidation during Human Pluripotent Stem Cell Lineage Specification." *Cell Cycle (Georgetown, Tex.)* 15 (18): 2464–75.

Woodward, Anna M., Thomas Göhler, M. Gloria Luciani, Maren Oehlmann, Xinquan Ge, Anton Gartner, Dean A. Jackson, and J. Julian Blow. 2006. "Excess Mcm2-7 License Dormant Origins of Replication That Can Be Used under Conditions of Replicative Stress." *The Journal of Cell Biology* 173 (5): 673–83.

Wu, Rentian, Zhiquan Wang, Honglian Zhang, Haiyun Gan, and Zhiguo Zhang. 2016. "H3K9me3 Demethylase Kdm4d Facilitates the Formation of Pre-Initiative Complex and Regulates DNA Replication." *Nucleic Acids Research*, September.

Wu, Shumin, and Judd C. Rice. 2011. "A New Regulator of the Cell Cycle: The PR-Set7 Histone Methyltransferase." *Cell Cycle (Georgetown, Tex.)* 10 (1): 68–72.

Wu, Shumin, Weiping Wang, Xiangduo Kong, Lauren M. Congdon, Kyoko Yokomori, Marc W. Kirschner, and Judd C. Rice. 2010. "Dynamic Regulation of the PR-

Set7 Histone Methyltransferase Is Required for Normal Cell Cycle Progression." *Genes & Development* 24 (22): 2531–42.

Yates, J. L., and N. Guan. 1991. "Epstein-Barr Virus-Derived Plasmids Replicate Only Once per Cell Cycle and Are Not Amplified after Entry into Cells." *Journal of Virology* 65 (1): 483–88.

Yates, John L. 1996. "Epstein–Barr Virus DNA Replication." In *DNA Replication in Eukaryotic Cells*, 751–774. Cold Spring Harbor Laboratory.

Yeeles, Joseph T. P., Tom D. Deegan, Agnieszka Janska, Anne Early, and John F. X. Diffley. 2015. "Regulated Eukaryotic DNA Replication Origin Firing with Purified Proteins." *Nature* 519 (7544): 431–35.

Yekezare, Mona, Belén Gómez-González, and John F. X. Diffley. 2013. "Controlling DNA Replication Origins in Response to DNA Damage – Inhibit Globally, Activate Locally." *J Cell Sci* 126 (6): 1297–1306.

Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, *et al.* 2008. "Model-Based Analysis of ChIP-Seq (MACS)." *Genome Biology* 9 (9): R137.

TABLE OF FIGURES

TABLE OF TABLES

ABBREVIATIONS

4-OHT	4-Hydroxytamoxifen
ARS	Autonomously replicating sequence
AS	Ascending segment
bp	Base pairs
BrdU	5-Bromo-2'-deoxyuridine
BSA	Bovine serum albumine
Cdc6	Cell division cycle 6
CDK	Cyclin dependent kinase
Cdt1	Cdc10 dependent transcript 1
ChIP(-seq)	Chromatin Immunoprecipitation (followed by sequencing)
CldU	5-Chloro-2'-deoxyuridine
CMG	Cdc45, MCM, Gins complex
CRL4	Cullin4A-ring ubiquitin ligase
CTCF	CCCTC-binding factor
DDK	Dbf4 Dependent Kinase
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphates
DOC	Deoxycholate
DS	*Depending on context: replication termination:* Descending segment
DS	*Depending on context: oriP:* Dyad symmetry element
EBV	Epstein-Barr virus
EdU	5-ethynyl-2'-deoxyuridine
FA	Formaldehyde
FACS	Fluorescence activated cell sorting
FR	Family of repeats
G1	Gap 1 phase during the cell cycle
G2	Gap 2 phase during the cell cycle
G4	G-quadruplex
H4K20me	Histone 4 lysine 20 methylation
HBOI	Histone acetyltransferase binding to ORC

HEK293 cells	Human embryonic kidney 293 cells
hES cells	Human embryonic stem cells
Hi-C	Chromosome Conformation Capture coupled to high-throughput sequencing
HP1	Heterochromatin protein 1
IdU	5-Iodo-2′-deoxyuridine
IGB	Integrated Genome Browser
kb	Kilobases
kDa	Kilodalton
LAD	Lamin associating domain
Mb	Megabases
Mcm2-7	Mini-chromosome maintenance proteins 2-7
MEF	Mouse embryonic fibroblast
MNase	Micrococcal nuclease
NDR	Nucleosome depleted region
NP-40	Ninidet P-40
OK-seq	Okazaki fragment sequencing
ORC	Origin recognition complex
ORCA	Origin recognition complex-associated
oriP	Epstein-Barr virus latent replication origin
PCR	Polymerase chain reaction
PI	Propidium iodide
pre-RC	Pre-replication complex
repl	Replicate
RNA	Ribonucleic acid
S	Synthesis-phase during cell cycle
S. cerevisiae	*Saccharomyces cerevisiae*
SDS	Sodiumdodecylsulfate
SEM	Standard error of the mean
SNS	Short nascent strands
TAD	Topological associating domain
TSS	Transcription start site
UAS	upstream activating sequence

RESUME SUBSTANTIEL

INTRODUCTION

La réplication d'ADN réfère au processus de duplication précise de l'information génétique afin d'assurer la transmission fidèle du génome pendant chaque division cellulaire. De ce fait, la réplication d'ADN est initiée une seule fois par cycle cellulaire. Une dérégulation de ce processus provoque de l'instabilité génétique, ce qui peut aboutir à des pathologies diverses, comme le cancer ou d'autres troubles génétiques. Les sites de l'initiation de la réplication d'ADN s'appellent « origines de réplication ». Sur le génome humain, entre 30000 et 50000 origines de réplication sont activées pendant chaque cycle cellulaire afin d'assurer la duplication complète du génome. Il y a un excès d'origines de réplication possible, de ce fait, uniquement une partie d'entre elles sont utilisées à chaque cycle réplicatif. Pour activer une origine, le chargement correct du complexe de pré-réplication (pre-RC) est requis. Le pre-RC consiste en deux sous-complexes majeurs : l' « origin recognition complex » (ORC) qui interagit directement avec l'ADN et est nécessaire pour recruter le second sous-complexe, les hélicases Mcm2-7, qui sont responsables de l'initiation de la réplication. Le pre-RC est uniquement formé pendant la phase G1 du cycle cellulaire, tant que l'activité de la kinase cycline dépendante (CDK) est basse. Avec l'activation de la CDK et de la DDK (kinase dépendante de DBF4) pendant la transition de G1/S, plusieurs protéines additionnelles se lient au pre-RC afin d'activer les hélicases Mcm2-7. En conséquence, le double hexamère Mcm2-7 se dissocie en deux hexamères actifs qui génèrent deux réplisomes constitués de plus de 150 protéines chacun qui assurent la réplication bidirectionnelle du génome.

Le choix de l'activation d'une origine est principalement stochastique. Le fait que seulement une partie des pre-RCs formés est activée au début de la phase S et que chaque cellule utilise une cohorte différente d'origines, a abouti à la définition du terme « efficacité de l'origine de réplication » qui représente la fréquence à laquelle une origine spécifique est activée dans une cellule donnée pendant un cycle cellulaire donné.

Recemment, plusieurs caractéristiques influencant l'efficacité de la formation et de l'activation d'une origine ont été identifié. Entre autre, la chromatine joue un rôle important dans le choix des origines de réplication. Il a été démontré que non seulement le positionnement des nucléosomes mais aussi les modifications des histones influencent la position et l'efficacité des origines de réplication.

La méthylation de l'histone H4 sur la lysine K20 (H4K20me) est notamment un candidat prometteur jouant un rôle dans la régulation de la réplication d'ADN pour des raisons multiples : i) l'expression de la transférase de monométhyle de H4K20, PR-Set7 (ou Set8, SetD8, KMT5A) est régulée au cours du cycle cellulaire, avec une concentration basse en G1, une absence complète en phase S, puis son expression augmente en G2 et atteint son maximum en mitose. ii) La stabilisation de l'expression de PR-Set7 pendant tout le cycle cellulaire induit la ré-réplication d'ADN, indiquant que la réplication a lieu plusieurs fois par cycle cellulaire. Une mutation additionnelle dans le domaine de méthylation de la protéine réverse ce phénotype, indiquant que la méthylation de H4K20 est un facteur régulateur de la réplication. iii) L'absence complète de PR-Set7 perturbe la progression cellulaire et est létale aux stages précoces du développement d'embryons de souris. PR-Set7 est la seule enzyme connue capable de catalyser la monométhylation de H4K20. Les méthylations additionnelles (di- et tri-méthylation) de H4K20 sont catalysées par les enzymes Suv4-20h1/h2. L'absence de Suv4-20h1/h2 provoque cependant un phénotype moins sévère que l'absence de PR-Set7. En conséquence, le rôle exact de PR-Set7 et H4K20me1 ou H4K20me3 dans la formation du pre-RC et son activation reste à découvrir.

Il est techniquement possible de cartographier (« mapping ») les évènements d'initiation de la réplication sur le génome humain par des approches différentes. A part le peignage d'ADN qui permet d'identifier des origines de réplication sur des molécules uniques, les méthodes les plus connues sont le séquençage des brins naissants (SNS-seq) et le séquençage des fragments d'Okazaki (OK-seq). La méthode du SNS-seq est basée sur l'isolation et le séquençage des brins précoces naissants des fourches de réplication actives. L'application d'une profondeur de séquençage élevée a révélé plus de 200000 origines de réplication actives dans une population de cellules asynchrones, représentantes des cohortes différentes d'origines potentielles utilisées par cellule. La technique de SNS-seq a été utilisée avec des cellules de variétés différentes, incluant les cellules souches embryonnaires humaines. Contrairement à cette méthode qui révèle une position précise de toutes les origines de réplication possible, le OK-seq sert à l'identification des larges zones *de préférence* d'activation et de terminaison de la réplication. En conséquence, une large zone d'initiation représente une zone préférentielle d'initiation de réplication, avec des sites d'initiation multiples mais seulement une origine de réplication active par cellule. OK-seq est une méthode plus récente et a été appliquée aux cellules cancéreuses HeLa et la lignée cellulaire de lymphome de Burkitt Raji infectée par le virus d’Epstein-Barr (EBV).

Les deux techniques de SNS-seq et OK-seq servent à identifier les origines activées dans une population de cellules. Pour identifier les positions *potentielles* d’activation de la réplication, il faut définir les positions du pre-RC dans les cellules. Les études génomiques par immuno-précipitation de la chromatine et le séquençage à haut débit (ChIP-seq) des pre-RCs peuvent être utilisé pour répondre à cette question. Aujourd’hui, les études de ChIP-seq des pre-RCs sont rares et jusqu’à maintenant seulement disponibles pour ORC. Etant donné que Mcm2-7 migre potentiellement de son site de chargement initial, il est crucial d'obtenir des informations sur le positionnement des Mcm2-7 pour la compréhension complète de la régulation de la réplication.

Des ChIP-seq de ORC et Mcm2-7 ont été réalisés sur le génome de EBV dans mon laboratoire hôte. L’EBV infecte les cellules B humaines et établit une infection latente persistante. Pendant l’infection latente, le génome de l’EBV est répliqué par la machinerie de réplication de l’hôte. La réalisation de ChIP ciblant les composants du pre-RC Orc2 et Mcm3, l’isolation parallèle de SNS et la digestion de la chromatine par la nucléase micrococcal (MNase) dans les cellules Raji, suivie par l’hybridation de l’ADN isolé sur une puce d’ADN désignée, a permis une analyse extensive de la relation entre la formation des pre-RCs, leur activation et le rôle du positionnement des nucléosomes dans ce contexte. Ces expériences fournissent un aperçu pour notre compréhension de l’organisation des origines de réplication dans des cellules humaines. Elles ont également montré que la technique du ChIP des composants du pre-RC est réalisable dans des cellules humaines et ont sensibilisé à la nécessité de contrôles adaptés. La prochaine étape sera la réalisation d’analyses similaires sur le génome humain entier afin de répondre à la question de la régulation des pre-RCs sur le génome humain.

BUT DE LA THESE

La régulation de l’initiation de la réplication a déjà été étudiée extensivement et cela a abouti à l’identification de plusieurs facteurs impliqués. Cependant, pour une compréhension complète de la régulation de la réplication, il nous faut connaitre la combinaison des informations sur l’initiation de la réplication et celles du positionnement du pre-RC. Notamment, le positionnement des Mcm2-7 – n’étant pas encore analysé jusqu’à présent – pourrait permettre de comprendre les divergences entre le ChIP-seq des ORC actuels et les données sur l’initiation de la réplication.

Le but de ma thèse sera d’utiliser la technique du ChIP-seq du pre-RC sur les génomes des cellules Rajis en ciblant les deux sous-unités majeures, ORC et Mcm2-7. Contenant le génome de l’EBV comme référence interne, les cellules Raji présentent un contrôle parfait de la qualité des résultats ChIP-seq. Une fois établi, ChIP-seq sera aussi utilisé dans les cellules souches embryonnaires humaines (cellules hES), en collaboration avec mon co-directeur de thèse Dr. Jean-Marc Lemaitre. La comparaison des informations sur le positionnement des pre-RCs avec l'initiation de la réplication, ainsi que l'identification des caractéristiques qui déterminent ce positionnement vont contribuer à la compréhension fondamentale de la relation entre la formation des pre-RCs et leur activation. De plus, la position des pre-RCs sera comparé avec certaines des caractéristiques de la chromatine, comme les modifications des histones. Il y a des preuves fonctionnelles que la méthylation de H4K20 est impliquée dans la régulation de la réplication de l'ADN, mais il nous manque la compréhension au niveau moléculaire. En conséquence, le ChIP-seq de H4K20me1 et –me3 sera aussi effectué, afin de conclure directement sur leurs implications possibles dans la régulation de la réplication d'ADN. Ces candidats seront validés en utilisant un système de plasmide basé sur l'EBV, afin de confirmer fonctionnellement la relation entre le méthylation de H4K20 et la réplication.

RESULTATS

OPTIMISATION, REALISATION ET VALIDATION DES CHIP-SEQ DU PRE-RC DANS LES CELLULES RAJI

Afin de répondre aux questions mentionnées auparavant, j'ai commencé par établir les expériences de ChIP-seq en cellules Raji. Pour l'optimisation du protocole du ChIP, afin d'assurer un maximum de rigueur et de spécificité, plusieurs considérations ont été pris en compte : i) une synchronisation des cellules en G1 permet d'effectuer le ChIP sur une population de cellules où le pre-RC est formé. Dans les cellules Raji, il est possible de fractionner les cellules à différents stades du cycle cellulaire selon leur taille. Ceci n'étant pas possible pour les cellules hES, ce ChIP fût conduit sur une population de cellules non-synchrone. ii) Une optimisation du temps de fixation des cellules (« cross-link ») ainsi qu'une sonication douce permet la préservation du complexe. iii) L'utilisation de différents anticorps pour cibler le même complexe consolide les résultats obtenus. J'ai décidé de cibler deux protéines de l'ORC : Orc2 et Orc3, ainsi que deux protéines de Mcm2-7 : Mcm3 et Mcm7. iv) L'augmentation de la spécificité des anticorps par l'optimisation des conditions des tampons est cruciale pour un résultat optimal.

L'utilisation des cellules Raji présente l'avantage de permettre le contrôle de la qualité des ChIPs ciblant les composants différents du pre-RC sur les régions connues pour être enrichies avec ces protéines par de la PCR quantitative en temps réelle (qPCR). Le locus choisi sur le génome de l'EBV est le « dyad symmetry element » (DS). La protéine virale, EBNA1, guide ORC à cette position, suivi par un recrutement des Mcm2-7, dépendant du cycle cellulaire. En effectuant ces contrôles par qPCR, ORC était présent dans les deux stades du cycle cellulaire choisis : G1 et S/G2. En revanche, Mcm3 et Mcm7 étaient plus enrichis en G1, ce qui valide d'une part la qualité des ChIPs et d'autre part le fractionnement des stades du cycle cellulaire. Des duplicates des ChIPs de Orc2 et Orc3, et trois exemplaires de Mcm3 et

Mcm7 en G1 et S/G2 ont été séquencés avec environ 70 millions de « reads » de profondeur. En parallèle, des duplicates des ChIPs de H4K20me1 et -me3 aux mêmes stades du cycle cellulaire ont été séquencés avec une profondeur de 45 millions de « reads ».

Afin de valider le ChIP et de choisir un programme de détection de pics, les « reads » du séquençage ont tout d'abord été alignés avec le génome de l'EBV. En comparant le résultat de trois différents programmes de détection de pics (MACS2, HOMER et T-PIC), avec des positions du pre-RC déterminées auparavant, j'ai établi que le programme T-PIC permettait de trouver les pics pertinents d'une manière sensible et fiable.

REALISATION CHIP-SEQ DE PRE-RC DANS DES CELLULES HES

L'analyse génomique des positions des composants du pre-RC dans les cellules Raji a également permis de déterminer des positions dans le génome qui permettent la validation des ChIPs du pre-RC par qPCR dans d'autres types cellulaires. Les ChIPs du pre-RC dans les cellules hES ont été validés sur l'origine de réplication connue Mcm4 et sur un locus identifié pour être enrichi dans les cellules Raji, BANF1. Après le séquençage, un premier aperçu des données a été obtenu. En collaboration avec Anissa Zouaoui, bioinformaticienne dans le laboratoire du Dr. Jean-Marc Lemaitre, l'identification de pics a été effectuée avec MACS2 et son paramètre pour déterminer des larges pics. Cette analyse nous a permis de déterminer que les composants de Mcm2-7 étaient beaucoup plus distribués que ceux de ORC. Par contre, ORC co-localisait plus avec les positions de la réplication active (obtenu par le SNS-seq), alors que les Mcm2-7 co-localisait à peine. Finalement, pour obtenir une meilleure idée des relations entre pre-RC et initiation de la réplication, il nous fallait une analyse plus précise et extensive. Dans ce but, nous avons utilisé le programme de détection de pics T-PIC (précédemment validé sur le génome de l'EBV) et analysé comment la couverture des « reads » est en corrélation avec les SNS, des positions des modifications de histones et d'autres éléments de la chromatine.

LES COMPOSANTS DU PRE-RC SONT ENRICHIS DANS LES ZONES DE LA REPLICATION ACTIVE

Afin de valider les données du ChIP-seq des composants du pre-RC dans les cellules Raji, nous les avons corrélés avec les zones d'activation et de terminaison de la réplication. Ces zones ont été déterminé par la méthode du OK-seq. Le séquençage des fragments d'Okazaki permet la différentiation en brins Watson et Crick. Grâce à cette différentiation, il est possible de calculer une « directionalité de la fourche de réplication », ce qui résulte en un profil de séries de segments ascendants (AS) et descendants (DS) de différentes tailles et pentes. Les AS représentent des zones préférentielles d'initiation de la réplication, alors que les DS sont des zones préférentielles de terminaison de réplication. L'amplitude y reflète l'efficacité de l'initiation ou de la terminaison.

En calculant la densité des pics de ORC ou Mcm2-7 dans les AS ou DS, j'ai montré que la densité des protéines du pre-RC dans des zones d'initiation de réplication est 2,5 fois plus élevée que dans les zones de terminaison.

Une autre façon d'analyser les données de séquençage est de calculer la couverture des « reads » sur une superposition soit d'AS, soit de DS. Cette méthode a l'avantage d'être indépendante de tout algorithme de détection de pics et permet l'obtention d'une tendance globale de la situation moyenne de tous les sites simultanément. La couverture moyenne de chaque membre du pre-RC était enrichie dans les zones d'initiation de réplication, et défavorisée dans les zones de terminaison de réplication. Ces résultats valident mon approche de ChIP-seq des protéines du pre-RC. De plus, la déplétion évidente des membres du pre-RC dans les zones de terminaison était inattendue et nécessite plus d'études afin de comprendre exactement les mécanismes impliqués.

LA POSITION DES MEMBRES DU PRE-RC EST PLUS STABLE DANS LES REGIONS DE REGULATION ACTIVE DE LA TRANSCRIPTION

Quelque ce soit la méthode d'analyse des origines de réplication, la réplication active est souvent liée à la transcription active. En conséquence, j'ai analysé la corrélation entre les membres du pre-RC et de la transcription active.

J'ai calculé la couverture moyenne des « reads » sur les sites d'initiation de la transcription (TSSs) actives déterminées dans les cellules Raji. La couverture était élevée sur les TSSs actives et non sur les TSSs inactives pour toutes les protéines du pre-RC, alors que ORC était légèrement plus élevé que Mcm2-7. Une analyse de détermination de pics montrait que les pics trouvés en proximité des TSSs actives sont plutôt forts et bien définis.

Une étude de la corrélation des membres du pre-RC avec des modifications des histones associées à la transcription active, H3K4me3 et H3K36me3, montrait que le pre-RC est localisé de façon préférentielle en association avec H3K4me3, une modification connue pour déterminer les promoteurs actifs. H3K36me3 est plutôt retrouvé dans le corps de gènes activement transcrits et ces positions ne sont pas enrichies du tout par les protéines du pre-RC.

LES HELICASES DU MCM2-7 SE RAPPROCHENT DE ORC EN S/G2

La liaison du pre-RC à l'ADN est limitée à la phase G1 du cycle cellulaire. En conséquence, pour un contrôle négatif de mes ChIP-seq, j'ai effectué les mêmes ChIPs dans la phase S/G2 du cycle cellulaire. Présumant que le pre-RC se dissocie de l'ADN soit à cause de leur activité de réplication, soit passivement, je me suis attendu à une diminution des membres du pre-RC aux positions identifiées auparavant.

Contrairement aux attentes, le nombre de positions des protéines du pre-RC, défini par les programmes de détection de pics n'ont pas changé notablement. De plus, la majorité des positions détectées coïncide avec les positions identifiées en G1. Par contre, l'analyse de couverture moyenne autour des zones d'initiation de la réplication (AS) et les zones de terminaison de la réplication (DS) ont montré une diminution de Mcm3 et Mcm7 autour des AS par rapport à G1 ainsi qu'une égalisation de la couverture autour des DS. Contrairement à cette observation, la couverture moyenne des membres du pre-RC autour des TSSs est encore plus élevée en S/G2 qu'en G1, identiquement à la couverture sur les positions de H3K4me3. De plus, l'analyse de détection de pics a montré que plus de pics pour Mcm2-7 étaient identifiés par le programme, ce qui indique un profil propagé de Mcm2-7 en G1 et un profil plus défini en S/G2. Ces observations indiquaient surtout que le positionnement de Mcm2-7 changeait en S/G2. Une analyse de la similarité des positions de ORC et Mcm2-7 en G1 et S/G2 montrait que les positions de Mcm2-7 se sont rapprochées de ORC en S/G2. L'explication possible de ce phénomène est que les hélicases Mcm2-7 sont initialement chargées sur des sites précis, déterminés par ORC, suivi par une migration de Mcm2-7 et éventuellement propagé par la transcription active.

ORC S'ASSOCIE A H4K20ME3 DANS L'HETEROCHROMATINE

Une question initiale de ma thèse était le rôle de la méthylation de H4K20 dans la régulation de la réplication. Dans ce but, des ChIPs additionnels ciblant soit H4K20me1 ou H4K20me3 ont été réalisés dans les mêmes conditions. Alors que H4K20me1 est une modification d'histone euchromatique, H4K20me3 fait partie des modifications des histones définissant l'hétérochromatine. Une analyse de la couverture moyenne des protéines du pre-RC aux positions soit de H4K20me1, soit de H4K20me3 a montré que le pre-RC n'est pas lié à H4K20me1, mais surtout que ORC est majoritairement enrichi à des positions proches de H4K20me3. Etant donné que ORC a aussi potentiellement un rôle dans la régulation de l'hétérochromatine, il fallait fonctionnellement tester le rôle de l'association de ORC avec H4K20me3.

H4K20ME3 EST NECESSAIRE POUR LA FORMATION DU PRE-RC ET SON ACTIVATION DANS L'HETEROCHROMATINE

En appliquant un test fonctionnel utilisant des plasmides autonomes basés sur l'EBV, il est possible de valider fonctionnellement le rôle d'ORC et H4K20me3. La stratégie étant de spécifiquement induire la tri-méthylation de H4K20 sur un locus du plasmide et de vérifier si cette induction est suivie par la formation du pre-RC sur ce locus et par une activation de la réplication du plasmide. En effet, l'induction de H4K20me1 sur le plasmide en guidant PR-Set7 sur le locus spécifique est suivie par une conversion en H4K20me3 par les enzymes Suv4-20h1/h2, le recrutement du pre-RC et l'activation de la réplication du plasmide. Pour déterminer si la conversion en H4K20me3 est nécessaire pour la réplication, l'utilisation d'un inhibiteur spécifique des enzymes Suv4-20h1/h2 empêchait la conversion de H4K20me1 en H4K20me3. En utilisant cet inhibiteur, la formation du pre-RC sur le plasmide est inhibée ainsi que sa réplication. Ces résultats montrent que H4K20me3 est en effet nécessaire pour la régulation de la réplication dans l'hétérochromatine. De plus, des expériences génomiques sur le timing de la réplication ont révélé des régions spécifiques dans l'hétérochromatine, qui dépendent de H4K20me3, alors que la réplication dans d'autres régions est potentiellement régulé différemment.

CONCLUSION

En conclusion, cette étude propose deux mécanismes différents de la régulation de l'assemblage du pre-RC dépendants de l'environnement de la chromatine. Dans la chromatine active, l'accessibilité de la chromatine joue un rôle très important pour le positionnement aléatoire d'ORC, en connexion avec la régulation de l'activité de la transcription. Le recrutement de multiples Mcm2-7 est possible et les Mcm2-7 sont probablement propagés par la machinerie de la transcription active dans les régions avoisinantes.

Dans l'hétérochromatine, une liaison aléatoire d'ORC est impossible à cause du compactage de la chromatine. En conséquence, une stabilisation d'ORC est nécessaire afin d'assurer la réplication correcte dans les domaines hétérochromatiques. Cette stabilisation dans l'hétérochromatine est entre autre accompli par H4K20me3. Les analyses génomiques ont démontré une association spécifique des sous-unités d'ORC avec H4K20me3, restreint dans des domaines définis par H4K20me3. Les analyses fonctionnelles ont précisé que H4K20me3 est nécessaire pour la formation et activation correcte du pre-RC.

L'ensemble de cette étude constitue la première analyse génomique par ChIP-seq de plusieurs composants du pre-RC, notamment car Mcm2-7 n'avait pas encore été analysé jusqu'à maintenant. Cette étude situe Mcm2-7 comme étant le déterminant majeur de l'initiation de la réplication dans l'euchromatine, alors que ORC représente le régulateur majeur dans l'hétérochromatine. De plus, elle fournit des preuves directes que la régulation de la formation du pre-RC dans l'hétérochromatine est dépendant de H4K20me3. Ces résultats ajoutent une pièce au puzzle de la régulation de la réplication humaine. Alors que quelques pièces déjà existantes exigent éventuellement une réévaluation, des investigations futures révèleront certainement d'autres éléments manquants afin de compléter l'image.

ACKNOWLEDGEMENTS

I would like to take the opportunity to thank first of all Dr. Aloys Schepers for his supervision and helpful discussions during my four years of PhD. He gave me plenty of rope without depriving me of his encouragement and even supported my sometimes rather uncomfortable ideas and wishes. Thank you, I am very grateful for that!

Many thanks also to my second PhD supervisor Dr. Jean-Marc Lemaitre, who provided me with everything I needed in his laboratory. He also made my participation at the CSH congress possible, an unforgettable experience.

I am also very grateful to our collaboration partner Dr. Eric Julien, not only for a great cooperation, but also for his commitment during congress preparations.

My sincere thanks also go to Dr. Axel Imhof and Dr. Gunnar Schotta as members of my thesis advisory committee for their helpful comments and encouragement during my PhD.

For his introduction in bioinformatics, I would like to thank Dr. Tobias Straub. He not only personally guided me through my inexperienced beginnings, but also patiently answered all of my questions – mostly twice. I also thank Anissa Zouaoui, who took over the bioinformatic analyses in hES cells.

I would also like to mention all of my collaborators who willingly shared their unpublished data with me: Alexander Buschle and Wolfgang Hammerschmidt, Muhammad Ahmad Maqbool and Jean-Christophe Andrau, and Xia Wu and Olivier Hyrien.

Finally, I would like to thank all current and past group members, especially Kris and Petra for their implication in my work, and the Schepers Racing Team (Steffi, Julia and Sarah) for a great working atmosphere, Sushi, ice battles and Disney songs.