Chromosome Research (2010) 18:79–89
DOI 10.1007/s10577-009-9094-2

Genome-wide approaches to determining origin distribution

Jean-Charles Cadoret · Marie-Noëlle Prioleau

Published online: 18 November 2009

Abstract Genome integrity depends upon a highly co-ordinated process that ensures the exact duplication of the genome at each cell cycle. Genomic mapping of DNA replication starting points in mammals, known as origins of replication, is an important step towards our understanding of how this essential mechanism is regulated throughout complex genomes. Two recent studies carried out in both human and mouse cells have revealed a strong association between replication origins and transcriptional regulatory elements. This strong overlap raises the question of how gene deserts, also lacking replication origins, are properly replicated in conditions where replication is disrupted. It also provides valuable information forward the identification of key regulatory factors of DNA replication initiation. Here, we review what these large-scale mappings of replication origins have brought to our understanding of replication initiation and what are the future prospects.

Keywords DNA replication origin · Chromatin · Transcription · CpG island · Large-scale mapping

Responsible editors: Marie-Noëlle Prioleau and Dean Jackson.

J.-C. Cadoret · M.-N. Prioleau (✉)
Institut Jacques Monod, Centre National de la Recherche Scientifique, Université Paris 7,
Paris 75013, France
e-mail: prioleau.marie-noelle@ijm.univ-paris-diderot.fr

Abbreviations

BrdUTP	5-Bromo-2′-deoxyuridine-5′-triphosphate
bZIP	DNA binding proteins containing two sub-domains, the leucine zipper and the basic region
Cdc6	Cell division cycle 6 protein
Cdt1	Chromatin licensing and DNA replication factor 1
CGI	CpG islands are genomic regions that contain a high frequency of CpG sites
CpG	Cytosine and guanine separated by a phosphate
DHS	DNase I hypersensitive site
ENCODE	Refers to a consortium, the Encyclopedia of DNA Elements
FISH	Fluorescence in situ hybridisation
HB01	Histone acetyltransferase binding to ORC1
HoxA	Is one of the four human Hox genes cluster expressing transcription factors carrying an homeodomain
HS	Hypersensitive site
IgH	Immunoglobulin heavy chain
Mcm 2-7	The minichromosome maintenance 2-7 complex functions as the eukaryotic replicative DNA helicase
ODP	Origin decision point
ORC	Origin recognition complex
pre-RC	Pre-replicative complex
SNS	Short nascent strands

SP1	Sp1 transcription factor
TSS	Transcription start site

Introduction

Complex genomes have the difficult task of ensuring that exactly one copy of their genome is duplicated in each cell cycle. Many studies have tried to determine how checkpoints efficiently control replication fork progression to prevent under-replicated regions from reaching the point at which chromosomes segregate. However, little is known about where replication forks are initiated, although such information would facilitate the prediction of the most critical regions for duplication. Cells start to prepare for DNA replication several hours before S-phase actually begins, with the assembly of pre-replication complexes (pre-RCs) at origins of replication during telophase and early G1 phase. Pre-RC assembly, also referred to as "origin licensing", involves the recruitment of Mcm2-7 protein complexes by the initiator proteins origin recognition complex (ORC), cell division cycle 6 protein and chromatin licensing and DNA replication factor 1. During G1, a transition known as the origin decision point occurs, fixing the position of the active pre-RCs (Dimitrova and Gilbert 1999). It is unknown whether this transition involves the selection of a subset of pre-RCs from an excess of licensed starting points or whether pre-RCs can move along chromosomes during G1. Either way, the final result is the establishment of a defined spatial programme fixing the points at which DNA replication starts. The preparation of this programme is crucial for the maintenance of genome integrity, because the disruption of pre-RC formation leads to genome instability (Lengronne and Schwob 2002; Dominguez-Sola et al. 2007). An understanding of the way in which this spatial programme is regulated is therefore of the utmost importance. This review focuses on the analysis of replication origins in vertebrates and therefore discusses recent progress only for these organisms. It has long been known that the ORC complex of metazoans, unlike that of *Saccharomyces cerevisiae*, displays no sequence specificity (Vashee et al. 2003). Based on this observation, it has been concluded that this property allows organisms in which different cells have different fates to adapt the duplication of their genomes to the establishment of new patterns of gene expression.

Until recently, our understanding on this complex aspect in vertebrates was based solely on information collected for a few selected model loci, making it difficult to develop an overall picture of the rules governing origin specification (Aladjem 2007). For example, only about ten of the estimated 50–100,000 origins in humans have been mapped. It rapidly became clear that origin selection did not depend on an obvious consensus sequence, so it was considered too risky to define rules based on the small collection of origin sequences available, which might turn out to be exceptions. The flexibility of mechanisms involved in origin selection necessitates the collection of a large body of data before defining rules.

The release of the genome sequences of several species and the development of large-scale methods for the systematic annotation of specific features have opened up new possibilities for formulating and addressing biological questions. Genome-wide analyses of the replication timing programme in mammals are described in other reviews in this issue and will therefore be discussed only when directly relevant to the spatial programme. However, by contrast to what has been achieved for simple yeast genomes, timing analyses in mammals have not resulted in the precise mapping of replication origins. This provides another indication of the complexity of the spatio-temporal programme in mammals, confirming that most origins are not systematically active in all cell cycles. It was therefore important to develop a sensitive and stringent method for the large-scale study of replication origins. New methods based on tiling micro-arrays are particularly suitable for this difficult task. We summarise here recent progress towards understanding what makes an origin of DNA replication. We will begin by describing the technical approaches used and the systems chosen. We will then discuss the results obtained in these analyses and future prospects.

Methods for constructing maps of replication origins

One of the major challenges in genome projects is to understand the molecular mechanisms generated by the genomic sequence itself. The identification of

DNA replication origins by genome-wide mapping is one way in which we are meeting this challenge, but the identification of these origins in higher eukaryotes has proved difficult because there seems to be no clear consensus sequence in these organisms. By contrast, such a consensus sequence has been identified in *S. cerevisiae*. Furthermore, the various methods used to locate origins are complex, because replication bubbles are transient and are therefore extremely scarce in populations of asynchronous cells. Until recently, fewer than 50 origins have been identified in higher eukaryotes, mostly in well-characterised transcribed regions. None of these previous studies defined a model for replication or attempted to classify origins on the basis of their properties. Two recent studies attempted to improve our understanding of the mechanism of DNA replication by developing high-resolution maps of replication origins in human and mouse cells based on the hybridisation to DNA micro-arrays of short nascent strands (SNS) isolated with the λ-exonuclease (Cadoret et al. 2008; Sequeira-Mendes et al. 2009). These studies increased the number of known origins of replication in mammals by a factor of ten.

Several protocols for the identification of origins of replication have been identified. Nascent strands can be labelled with 5-bromo-2′-deoxyuridine-5′-triphosphate or radiolabelled nucleotides in the early S-phase of synchronous cell cultures, making it possible to detect the synthesis of new fragments. Origins can then be detected by hybridisation with cloned genomic DNA. There are two drawbacks to this method: (1) it cannot detect origins activated at other moments during S-phase and (2) the use of drugs to induce synchronisation may introduce artefacts into the normal replication programme.

A second method involves coupling DNA combing and fluorescence in situ hybridisation (FISH). This approach requires both a mastery of single-molecule techniques and a powerful, prolonged analysis of DNA fibres. This process is currently too slow for large-scale analyses (Anglana et al. 2003; Norio et al. 2005).

A third method is based on the trapping of bubble-shaped structures corresponding to the initiation of replication from a sample of restricted fragments of genomic DNA in an agarose gel. A library of trapped fragments can then be constructed and used as a probe in a genome-wide approach (Mesner et al. 2006). However, the accuracy and resolution of this method may fluctuate around a few kilobases, rendering this method unsuitable for detailed analyses of correlation with elements of the genome.

Finally, a stringent method based on the resistance of SNS to λ-exonuclease digestion has been developed for origin isolation and is suitable for high-throughput approaches (Gerbi and Bielinsky 1997). This method was used by the only two groups to have published coherent data concerning the location of origins in a genome to date.

Cadoret et al. (2008) worked with HeLa-S3 cells and extracted SNS for hybridisation with DNA micro-arrays covering the Encyclopedia of DNA Elements (ENCODE) regions (Fig. 1a). The first phase of the ENCODE project involved the reporting and analysis of data from different experiments carried out by various groups on a targeted 1% of the human genome (ENCODE 2004). By focusing on these genomic regions, Cadoret et al. (2008) were able to highlight correlations between the location of origins of replication and various annotations of elements in these regions, such as histone modifications. Only a small fraction of the genome was covered by this dataset (1%), but it should be possible to generalise the conclusions drawn from this analysis because ENCODE regions were 44 discrete regions selected as regions representative of the whole genome.

As small numbers of SNS can be recovered with the λ-exonuclease method, this laboratory chose to amplify the sample by T7-based DNA linear amplification, which results in only low levels of bias. The authors believe that they may have missed 5% to 10% of the true positives due to the stringency of their method, but this approach seems to be specific and sensitive enough to detect most of the origins active in at least 10% of cell cycles, regardless of the timing of their activation during S-phase. This study also demonstrated that a preparation of SNS not purified with λ-exonuclease described in a recent paper was not suitable for use in the high-throughput mapping of replication origins because the SNS in such preparations are greatly outnumbered by large amounts of fragmented genomic DNA. It is therefore impossible to measure the abundance of genuine SNS. It has thus clearly been demonstrated that the identification of origins through SNS preparations requires λ-exonuclease digestion.

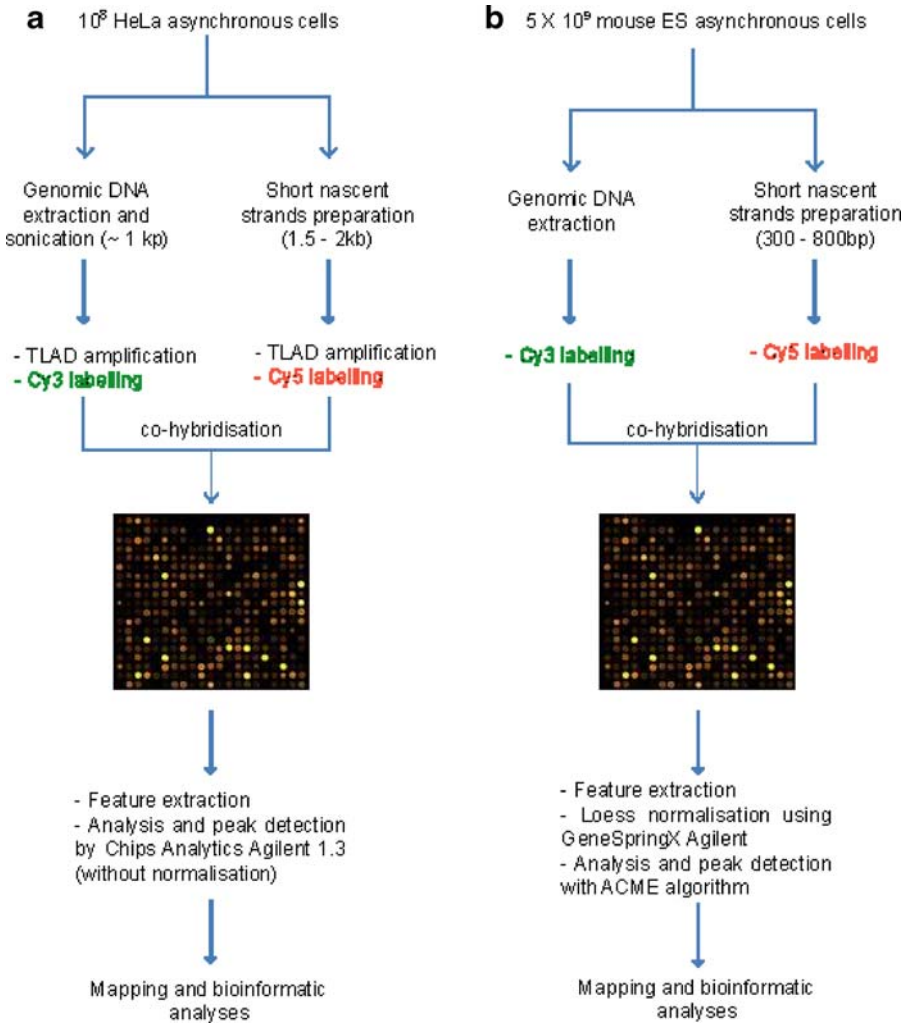

Fig. 1 Different strategies for the high-resolution mapping of DNA replication origins in two different mammalian genomes. **a** Short nascent strands (SNS), 1.5–2 kb, were isolated with the λ-exonuclease from 10^8 human HeLa cells. SNS were then amplified with the TLAD method including an indirect labelling with Cy5 (*red*). Genomic DNA was extracted, sonicated and then amplified and labelled with Cy3 (*green*) with the same method as the one used for SNS. The Cy5- and Cy3-labelled cRNAs were mixed in equal amount and co-hybridisated on agilent ENCODE ChiP-on-chip microarray. Feature extraction was done by Feature Extraction 9.1 software and analysed by Chip Analytics 1.3. **b** SNS, 300–800 bp, were isolated with λ-exonuclease from 5×10^9 mouse ES cells. Sample labelling, hybridisation and data extraction were performed according to standard procedures from Agilent Technologies

The second study, from the group of Gomez, described a similar approach to the mapping of replication origins in 0.4% of the genome of mouse embryonic stem cells (Fig. 1b.). The SNS were again isolated by λ-exonuclease digestion, providing a further demonstration of the efficacy and sensitivity of this method. However, unlike Cadoret et al. (2008), this group decided not to amplify the SNS for microarray hybridisation. Instead, the SNS were isolated from the 300–800 nt fraction obtained from 5×10^9 mouse embryonic stem cells. By contrast, Cadoret's extracts were obtained from the 1.5–2 kb fraction from 10^8 human HeLa cells. The study in mouse cells therefore had the advantages of providing a more quantitative vision of origin efficiency and a higher level of resolution concerning the position of the origin.

These two studies reported a large number of new origins, generating a general vision of the spatial programme operating in the genome and constituting a statistically relevant dataset for studies of the mechanism underlying the choice of replication origin. Data are now available from other laboratories, making it possible to check the correlation between origin position and position of other genomic elements, such as transcriptional regulatory elements, CpG islands (CGI) and promoter regions. Galaxy 2^{ENCODE}, a web application, can be used to manipulate genomic intervals in various ways, facilitating intersections, subtractions and merges (http://main.g2.bx.psu.edu/; Blankenberg et al. 2007). Other tools may have similar functions and uses for the analysis of genomic data, but the Galaxy 2^{ENCODE} application has the advantage of providing experimental biologists with an intuitive and easy-to-use interface.

The groups responsible for these two studies carried out thorough analyses. They began by investigating the distribution and density of origins within the genome. They then compared the positions of the origins detected with the positions of regions conserved throughout evolution. They then checked for the co-localisation of origins and markers of gene regulation or transcription initiation activity. These studies have thus led to significant progress in the characterisation of origins, but data collection is far from complete, as the origin sequences identified have not yet divulged the "secrets" of origin specification in the genome.

The genome contains large regions devoid of strong origins, constituting a risk for genome integrity

The studies in human HeLa and mouse ES cells both showed a high degree of correlation between origin density and GC (or gene) richness. This correlation is a consequence of the strong association between origins and annotated promoters (44% in mouse and 28% in humans) or more distal transcriptional regulatory elements. This observation is consistent with the long-standing correlation between early replication and GC-richness. Regions that are replicated early must contain replication origins, but they may have a few highly efficient origins or a large number of weak origins clustered into a dense array. Conversely, domains replicated later may make use of "passive replication" processes involving forks initiated at origins located in the surrounding regions. Indeed, a precise inspection of replication timing in mouse ES cells showed that, in many cases, there is a gradual change in the timing of replication over the length of a particular region (Farkash-Amar et al. 2008; Hiratani et al. 2008). There was a uniform fork direction in every region with such a gradual temporal programme studied. These findings strongly suggest that such regions represent large replicons. These studies therefore suggest that approximately 10% of the genome consists of large replicons, in which replication is initiated at a distant origin. We can therefore predict that at least 10% of the genome is devoid of efficient origins of replication. The extent of these origin-less regions is underestimated, because late-replicating regions devoid of origins and next to transition zones were not taken into account. The immunoglobulin heavy chain locus (*Igh*) in mouse provides a well-studied example of a transition region replicated by a single replication fork. Detailed analyses of replication timing coupled with evaluations of fork direction have shown that the 450-kb *Igh* region is bounded on one side by a domain that is replicated early and on the other by late-replicating region (Ermakova et al. 1999). Recent analyses of origin dynamics by single-molecule analyses of replicated DNA have confirmed this observation (Norio et al. 2005).

The replication mapping of ENCODE regions, which are representative of the whole human genome, confirmed predictions relating to replication timing. The 44 regions covering 30 Mb were found to include six regions of 500 kb devoid of strong sites of initiation, consistent with the average percentage of regions of replication timing transition found in mouse. Replication timing analyses of the centre of these regions showed them to be replicated in mid or late S-phase. They were therefore either transition zones or late-replicated domains in which forks progressed in a single direction away from a transition zone. Further confirmation of the existence of large replicons, greater than 1 Mb in size, was obtained by the pioneering experiments of Yrov and Liapunova in 1977, in which stretched DNA molecules were labelled with ^{3}H-thymidine and subjected to autoradiography (Yurov and Liapunova 1977). All the experiments carried out in different systems and with different methods suggest that a fraction of the genome lacks potential sites of replication initiation. So what has the genome-wide mapping of replication origins contributed to replication timing analyses and fibre analyses? Firstly, the unidirectional movement of a replication fork in replication timing analyses is not sufficient to prove the absence of origins of replication in a given region, as there is an alternative scenario in which origins are gradually activated by the replication fork. Origin mapping suggests that if this were actually the case, then the initiation of replication would be so dispersed that it would be impossible to detect any specific, focused site of replication initiation. Molecular combing or fibre spreading without FISH analyses, although providing a powerful vision of the variation of replicon size and organisation, cannot provide information about which regions contain dense clusters of origins or large

replicons. Large-scale studies are making it possible to determine the nature of these regions, which consist largely of AT-rich regions lacking annotated genes (also known as gene deserts).

One important question raised by the observation that some regions correspond to large replicons concerns the maintenance of genomic stability at these loci. It has been suggested that origin-less regions are the most sensitive to replicative stress, because there is likely to be a much lower probability of having a backup origin (licensed origin) between two converging collapsed replication forks than in origin-dense regions. Indeed, a survey of a few such boundaries on human chromosomes 11 and 21 showed these regions to be correlated with genes that are frequently disrupted in cancer (Watanabe et al. 2004). Two studies recently showed that an excess of Mcm2-7 loading over the number of replication origins during G1 is required for the activation of "dormant" (or "cryptic") origins in cases of replicative stress (Ge et al. 2007; Ibarra et al. 2008). DNA fibre analysis was used to monitor the mean fork spacing within clusters of origins. This study showed that decreasing the rate of DNA synthesis resulted in a decrease in replication fork spacing, from approximately 25 to 17 kb. This study analysed regions containing origin clusters and, therefore, does not specifically address the question of how regions normally lacking origins of replication behave. The recently developed SNS-on-chip should make it possible to determine whether "cryptic" origins are activated under replicative stress. One alternative possibility is that initiation is less focused in such conditions and therefore not detectable by this method. These regions could also be analysed by selecting a subset of 100–200 kb origin-less regions (at least ten such regions to optimise the number of positive fibres on FISH analysis) and applying a protocol comparing origin density in normal and replicative stress conditions. Finally, chromatin immunoprecipitation experiments with antibodies against Mcm sub-units could be used to determine whether Mcm2-7 complexes are loaded evenly throughout the genome in G1. If, as for the principal origins used, density was found to be correlated with GC and gene density, then AT-rich regions lacking origins would be clearly identified as zones in which genome integrity was at risk.

Origin density is not homogeneous and is not strictly related to replication timing

The range of origin densities in ENCODE regions was broad, with a minimum of no origin in 500 kb and a maximum of one origin every 11 kb (HoxA locus, 12 origins in 135 kb). As most origins have a firing efficiency below 50%, analyses based on studies of populations yield a cumulative picture of origin activity and therefore underestimate the accurate distance between origins on an individual DNA molecule. Only experiments based on single-molecule analyses can provide a true vision of inter-origin distances in a single cell. However, if we assume that most origins have similar efficiencies, such analyses show that the genome is replicated by extremely variable replicons. One region of particular interest is the HoxA locus, which has an origin density of one origin/11 kb, with high levels of SNS enrichment at sites of detected peaks (origins), similar to that at the canonical c-myc origin or at isolated origins found elsewhere. This results that, for some loci, several origins located close together may become active on the same DNA molecule at the same time. As previously suggested for origin-less regions, the fork density of a subset of regions with a high origin density could be analysed by molecular combing associated with FISH. The prediction is that these regions would be more likely to give figures containing short replicons arranged in tandem, with origin firing at approximately the same time. If existing, it is possible that dense clusters of synchronous replicons require a particular chromosomal organisation. The identification of such potential regions is now making it possible to test this hypothesis.

Over the last few years, there has been considerable discussion about the general existence of temporal programming of origin activation in eukaryotes (Rhind 2006). Early origins presumably correspond to efficient initiation, whereas late origins correspond to inefficient sites of initiation. The dense clusters of origins (three origins in 20 kb) identified in the ENCODE regions are small domains of highly efficient replication initiation (i.e. a high percentage of the cells displayed replication initiation in this cluster in the population analysed). An analysis of replication timing for ten dense clusters showed that half these clusters were not activated early in S-phase. Indeed, one of these clusters

initiated replication in mid-late S-phase. Isolated origins (one origin in a window of 200 kb) show also a wide range of replication timing. These results are difficult to reconcile with a model in which there is no specific temporal programme controlling the timing of origin activation. These examples illustrate the complexity of the replication programme in mammals and suggest, as put forward in several studies, that the chromosomal environment plays an important role in controlling origin firing. This issue could be specifically addressed in higher eukaryotes by transferring origin clusters to different chromosomal positions lacking strong sites of initiation and investigating the effects of this insertion on the temporal programme. Many origins overlap with transcriptional regulatory elements also known to be involved in chromatin structure regulation, so some clusters may carry not only information relating to strong origins but also *cis*-elements involved in the control of temporal firing. No matter how complex replication timing turns out to be, the identification of discrete sites of replication initiation and the precise mapping of replication timing are required for identification of the molecular mechanisms responsible for the firing of origins of replication.

Origin selection and transcription factors

Until recently, too few origins had been mapped in vertebrates to address the specific question of whether replication origins tend to be associated with transcriptional regulatory elements. Moreover, as the regions scrutinised tended to be regions surrounding coding genes, the reported distribution of origins close to gene promoters may result simply from a lack of investigation of gene-free regions. ENCODE regions were chosen by the consortium so that they are representative of the whole human genome and, therefore, studies along these regions should give a comprehensive vision of the genome activity. Large-scale studies of both gene-dense and gene-desert regions have shown that origin density and gene density are correlated in both human and mouse cells, reflecting the co-ordinated organisation of replication and transcription. This finding suggests that these two processes may have regulatory factors in common. The most striking result in both studies was the strong association between origins and CGI (50% of origins in human HeLa cells and mouse ES cells are located close to a CGI). This shows the conservation of origin specification mechanisms across species and cell types. Moreover, quantitative analyses in mice have shown that the origins associated with CGI promoters are the most efficient origins, consistent with competent pre-RC complex recruitment being strongly favoured at these sites. In the study carried out in human cells, the resolution of origin mapping is about 500 bp and, therefore, it does not give the information whether replication initiation is located precisely at transcriptional regulatory elements or just nearby. In mouse, where very SNS (300–800 bp) were used allowing a better resolution, a remarkable parallel organisation of replication initiation sites and transcription start sites (TSS) has been reported, suggesting that the transcriptional and replication machineries may make use of similar molecular mechanisms.

CGI are known to be bound by many transcription factors, and the most common regulatory motif identified in CGI is the binding site for Sp1 transcription factor. It is therefore difficult to identify the most important factors for origin selection. The most plausible hypothesis is that several combinations of a group of transcription factors lead to the efficient recruitment of pre-RC complexes (Fig. 2). This would account for the problems identifying a specific consensus sequence element. However, it is possible to identify specific transcription factors involved in this regulation. The use of regions explored by the ENCODE Consortium made it possible to compare origin positioning with the distribution of several features, including the binding of transcription factors. Chromatin immunoprecipitation on chip (ChIP-on-chip) data concerning transcription factor binding showed that the c-Jun and c-Fos transcription factors (forming the AP-1 complex) were significantly associated with 20% of the mapped origins and that this strong association could not be attributed solely to the coincidence of origins and CGI. These findings suggest that the AP-1 complex may be a key regulator of replication initiation. The reporting, by two different groups, of a role for the HBO1 histone acetyltransferase, a specific coactivator of the AP-1 subfamily of bZIP proteins, in the regulation of replication origin licensing supports this hypothesis (Iizuka et al. 2006; Miotto and Struhl 2008). These observations suggest that AP-1 recruits HBO1, in turn favouring the formation of a functional pre-RC complex.

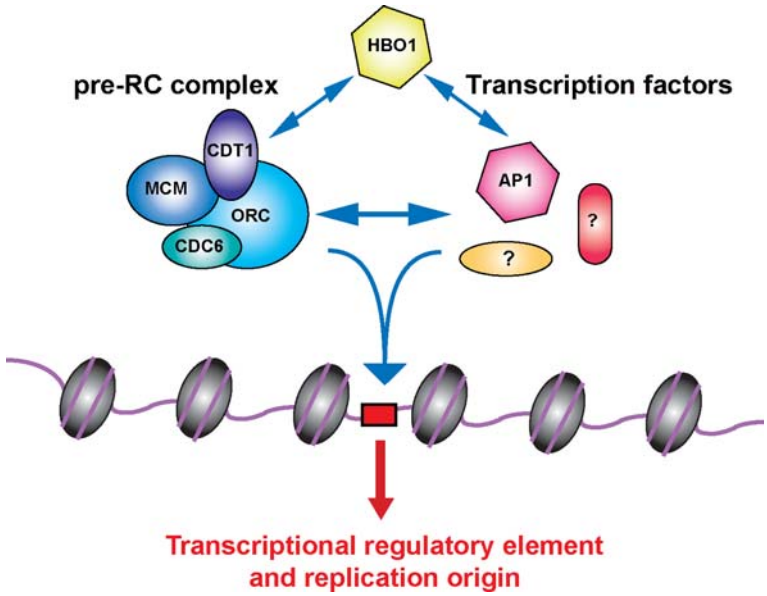

Fig. 2 Transcription factors control origin selection: large-scale mapping of replication origins in human and mouse cells showed that most mapped origins overlap with transcriptional regulatory elements suggesting a role of transcription factors in regulating origin selection. This could be achieved by either a direct or indirect recruitment of sub-units of the pre-replicative complex by transcription factors. The AP-1 complex found associated with 20% of origins mapped in human cells is proposed to be involved in origin selection. AP-1 specifically recruits the histone acetyltransferase HBO1. This factor has been shown to control origin licensing in agreement with the role of AP-1 in origin selection

The release of increasing amounts of data concerning transcription factor binding sites should also make it possible to investigate the association of other factors with replication origins. The c-*myc* proto-oncogene has also been identified as potentially important for origin activation, but the ChIP-on-chip data obtained by two groups for HeLa cells showed no significant overlap between the binding sites for this factor and replication origins (Cadoret et al., unpublished data). It remains unclear whether this finding reflects an inability to immunoprecipitate c-*myc* bound at origins or whether c-*myc* plays no direct role in origin selection in human cells. The large-scale mapping of replication origins could also facilitate the search for de novo motifs. The composite organisation of replication origins is a major obstacle to such studies, but at least we now have the possibility of exploring this alternative way of working.

Origin selection and chromatin structure

As no consensus sequence for origins has been identified, it was considered possible that chromatin modifications might play an important role in origin of replication selection. The most commonly described epigenetic mark is histone acetylation, which also serves as a major regulator of transcription. This mark, which is known to be associated with all active promoters, has been identified as a candidate regulator of origin specification. This hypothesis has been confirmed by only one observation in metazoans. Moreover, in this isolated case, origin activation controls amplification of the DNA at the *Drosophila* chorion locus and may therefore correspond to a very specific method for regulating the initiation of replication (Aggarwal and Calvi 2004). Chromatin hyperacetylation at this locus is critical for recruitment of the ORC complex and re-replication. Further support for this hypothesis is provided by the observation that, in *Xenopus* eggs, the binding of a transcription complex close to the *c-myc* promoter leads to local hyperacetylation and the induction of site-specific replication initiation (Danis et al. 2004). However, it was not possible in this study to determine whether histone acetylation or the binding of transcription factors was the key event in origin specification. Origin mapping to specific loci in mouse and

chicken cells showed that histone acetylation marks were not systematically present at efficient origins (Prioleau et al. 2003; Gomez and Brockdorff 2004; Dazy et al. 2006; Gregoire et al. 2006). In conclusion, the few examples tested have demonstrated that histone acetylation is not a prerequisite for the formation of a competent pre-RC.

Large-scale analyses have made it possible to determine, for a statistically relevant data set, whether active sites of replication initiation tend to be associated with hyperacetylated histones. Again, the ENCODE regions provide a very powerful system for study, because DNase I hypersensitive sites (DHS), histone H3 and H4 acetylation, histone H3K4 mono, di and trimethylation have all been mapped in HeLa cells (Koch et al. 2007; Xi et al. 2007). Consistent with the histone code hypothesis, multiple histone modifications act in a combinatorial manner to specify different chromatin states. H3K4me2, H3K4me3 and H3ac are highly correlated and are also found associated with DHS. These regions are markers of active promoters and are also significantly correlated with replication origins. Associations of this type seem to be an advantage but are not absolutely required, as approximately 45% of origins have neither histone modifications nor a hypersensitive site. Moreover, CGI harbouring H3K4me3 are not better substrates for origin specification than CGI lacking this marker. We cannot rule out the possibility that histone modifications known to be associated with active transcription are important for a subset of origins, but this study demonstrates that a large proportion of origins are not regulated by these canonical histone marks. As only a limited number of histone modification marks were assayed, it remains possible that origins lacking the classical markers of open chromatin have an unexplored modification. Alternatively, it is also possible that a very transient modification of histones is required during origin licensing, to provide the environment for pre-RC binding, and that studies in asynchronous cells are unable to detect such transient events. In any case, it is clear that the strong association between regulatory elements and origin positioning is not directly controlled in the same way as gene expression. The replication and transcriptional machineries have evolved to make use of similar factors, but with different optimal combinations. This probably reflects the differences in the constraints imposed on a strong promoter, which must generate a large number of transcripts in a short period of time, and those imposed on an origin of replication, which must fire once and only once at each cell cycle. Chromatin organisation may play a role in these differences, but it remains unclear which properties of chromatin are critical for origin activity.

The sequences of origins of replication are constrained, whereas their position with respect to gene orientation is not

We have previously described the non-random distribution of origins in mammalian genomes, with origin density being higher in gene-rich regions. This association is due to the strong overlap between origins of replication and transcriptional regulatory elements. These elements are known to be constrained during evolution and are therefore expected to show high levels of sequence conservation. The level of sequence conservation for origins of replication in humans was assessed by analysing the fraction of origins in genomic regions strongly conserved in mammals. As expected, proximity to a TSS and CGI increases the level of sequence conservation. However, origins display significantly higher levels of sequence conservation than would be expected on the basis of chance alone, regardless of their distance from a TSS or CGI (Necsulea et al. 2009). It therefore seems likely that specific sequence elements required for origin function and not solely for promoter function are conserved. This would suggest that there are specific *cis*-regulatory elements associated with origin function.

Another important question analysed with data collected by Cadoret et al. (2008) concerned the possible selection of replication start sites during evolution to prevent head-on collisions between replication and transcription polymerases. Such a "replication-related organisation" of the genome has been observed in bacteria and was recently suggested to apply to humans too, on the basis of *in silico* predictions of replication origins (Rocha 2004; Huvet et al. 2007). However, experimental data have provided no support for this hypothesis. There is no evidence for selective pressure to prevent collisions between the replication and transcription machineries (Necsulea et al. 2009).

Outlook

Identification of the molecular mechanisms responsible for origin selection in mammals has been hampered by the lack of a comprehensive knowledge of the origins of replication in the genome and by problems establishing a powerful genetic model system for accurately and easily detecting origin activity. The development of new methods of large-scale mapping on DNA micro-arrays has undoubtedly opened up new perspectives. The results obtained for very small portions of the human and mouse genomes (approximately 1%) strongly suggest a role of transcriptional functional elements, such as CGI and promoters, in the regulation of DNA replication. A combination of the increasingly commonly used deep-sequencing methods and powerful statistical analyses should make it possible to map origins for entire genomes. Moreover, the collection of data for different cell types and at different stages of differentiation should provide important insight into the coupling of DNA replication to the establishment of expression programmes. The complex connections between transcription and origin selection observed in ENCODE regions in HeLa cells suggest that it will be difficult to define clear rules. Such studies should provide us with information about the proportion of origins found in most cell types. If a large pool of "constitutive origins" were to be identified through such studies, it would facilitate the precise definition of what is shared, making it possible to identify critical *cis*-regulatory elements. The identification of "tissue-specific origins" would also provide valuable information about the key events involved in the activation of these origins, particularly for well-characterised loci. One example of well-known developmentally regulated origins of replication in mouse concerns the origins activated within the *Igh* locus. The activation of these origins is characterised by a change in replication timing in the chromosomal region, from late to early S. Regions undergoing a change in replication timing during differentiation should be considered likely to contain tissue-specific origins.

The power of genome-wide studies lies in their ability to test a broad range of data and to extract significant overlaps between a site of productive initiation and another property. This makes it possible to identify potentially critical events in the formation of a productive replication complex. However, improvements in our understanding of the efficient formation of replication complexes at active sites of replication initiation will require the development of both genome-wide maps for different cell types and efficient genetic tools for validating new hypotheses. The construction of origin maps and the identification of the transcriptional regulatory elements associated with them should greatly facilitate the establishment of new model systems. Finally, combining analyses of origin firing at specific loci by in vivo labelling and single-molecule analysis by DNA combing with whole genome approaches which cannot account for origin redundancy will ultimately give a vision which hews closely to the true complexity of origin firing in vertebrate genomes.

Acknowledgments J-C.C has a post-doctoral fellowship from the Agence Nationale pour la Recherche (ANR-08-BLAN-0080-01).

References

Aggarwal BD, Calvi BR (2004) Chromatin regulates origin activity in *Drosophila* follicle cells. Nature 430 (6997):372–376

Aladjem MI (2007) Replication in context: dynamic regulation of DNA replication patterns in metazoans. Nat Rev Genet 8(8):588–600

Anglana M, Apiou F, Bensimon A, Debatisse M (2003) Dynamics of DNA replication in mammalian somatic cells: nucleotide pool modulates origin choice and inter-origin spacing. Cell 114(3):385–394

Blankenberg D, Taylor J, Schenck I, He J, Zhang Y et al (2007) A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. Genome Res 17(6):960–964

Cadoret JC, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C et al (2008) Genome-wide studies highlight indirect links between human replication origins and gene regulation. Proc Natl Acad Sci U S A 105(41):15837–15842

Danis E, Brodolin K, Menut S, Maiorano D, Girard-Reydet C et al (2004) Specification of a DNA replication origin by a transcription complex. Nat Cell Biol 6(8):721–730

Dazy S, Gandrillon O, Hyrien O, Prioleau MN (2006) Broadening of DNA replication origin usage during metazoan cell differentiation. EMBO Rep 7(8):806–811

Dimitrova DS, Gilbert DM (1999) The spatial position and replication timing of chromosomal domains are both established in early G1 phase. Mol Cell 4(6):983–993

Dominguez-Sola D, Ying CY, Grandori C, Ruggiero L, Chen B et al (2007) Non-transcriptional control of DNA replication by c-Myc. Nature 448(7152):445–451

ENCODE pc (2004) The ENCODE (ENCyclopedia Of DNA elements) project. Science 306(5696):636–640

Ermakova OV, Nguyen LH, Little RD, Chevillard C, Riblet R et al (1999) Evidence that a single replication fork proceeds from early to late replicating domains in the IgH locus in a non-B cell line. Mol Cell 3(3):321–330

Farkash-Amar S, Lipson D, Polten A, Goren A, Helmstetter C et al (2008) Global organization of replication time zones of the mouse genome. Genome Res 18(10):1562–1570

Ge XQ, Jackson DA, Blow JJ (2007) Dormant origins licensed by excess Mcm2–7 are required for human cells to survive replicative stress. Genes Dev 21(24):3331–3341

Gerbi SA, Bielinsky AK (1997) Replication initiation point mapping. Methods 13(3):271–280

Gomez M, Brockdorff N (2004) Heterochromatin on the inactive X chromosome delays replication timing without affecting origin usage. Proc Natl Acad Sci U S A 101 (18):6923–6928

Gregoire D, Brodolin K, Mechali M (2006) HoxB domain induction silences DNA replication origins in the locus and specifies a single origin at its boundary. EMBO Rep 7 (8):812–816

Hiratani I, Ryba T, Itoh M, Yokochi T, Schwaiger M et al (2008) Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol 6 (10):e245

Huvet M, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y et al (2007) Human gene organization driven by the coordination of replication and transcription. Genome Res 17(9):1278–1285

Ibarra A, Schwob E, Mendez J (2008) Excess MCM proteins protect human cells from replicative stress by licensing backup origins of replication. Proc Natl Acad Sci U S A 105(26):8956–8961

Iizuka M, Matsui T, Takisawa H, Smith MM (2006) Regulation of replication licensing by acetyltransferase Hbo1. Mol Cell Biol 26(3):1098–1108

Koch CM, Andrews RM, Flicek P, Dillon SC, Karaoz U et al (2007) The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Res 17(6):691–707

Lengronne A, Schwob E (2002) The yeast CDK inhibitor Sic1 prevents genomic instability by promoting replication origin licensing in late G(1). Mol Cell 9(5):1067–1078

Mesner LD, Crawford EL, Hamlin JL (2006) Isolating apparently pure libraries of replication origins from complex genomes. Mol Cell 21(5):719–726

Miotto B, Struhl K (2008) HBO1 histone acetylase is a coactivator of the replication licensing factor Cdt1. Genes Dev 22(19):2633–2638

Necsulea A, Guillet C, Cadoret JC, Prioleau MN, Duret L (2009) The relationship between DNA replication and human genome organization. Mol Biol Evol 26(4):729–741

Norio P, Kosiyatrakul S, Yang Q, Guan Z, Brown NM et al (2005) Progressive activation of DNA replication initiation in large domains of the immunoglobulin heavy chain locus during B cell development. Mol Cell 20(4):575–587

Prioleau MN, Gendron MC, Hyrien O (2003) Replication of the chicken beta-globin locus: early-firing origins at the 5′ HS4 insulator and the rho- and betaA-globin genes show opposite epigenetic modifications. Mol Cell Biol 23 (10):3536–3549

Rhind N (2006) DNA replication timing: random thoughts about origin firing. Nat Cell Biol 8(12):1313–1316

Rocha EP (2004) The replication-related organization of bacterial genomes. Microbiology 150(Pt 6):1609–1627

Sequeira-Mendes J, Diaz-Uriarte R, Apedaile A, Huntley D, Brockdorff N et al (2009) Transcription initiation activity sets replication origin efficiency in mammalian cells. PLoS Genet 5(4):e1000446

Vashee S, Cvetic C, Lu W, Simancek P, Kelly TJ et al (2003) Sequence-independent DNA binding and replication initiation by the human origin recognition complex. Genes Dev 17(15):1894–1908

Watanabe Y, Ikemura T, Sugimura H (2004) Amplicons on human chromosome 11q are located in the early/late-switch regions of replication timing. Genomics 84(5):796–805

Xi H, Shulha HP, Lin JM, Vales TR, Fu Y et al (2007) Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. PLoS Genet 3(8):e136

Yurov YB, Liapunova NA (1977) The units of DNA replication in the mammalian chromosomes: evidence for a large size of replication units. Chromosoma 60(3):253–267