Initiation of DNA Replication in the Human Genome

Manuel S. Valenzuela

Department of Biochemistry and Cancer Biology, School of Medicine, Meharry Medical College, 1005 D.B. Todd Jr. Boulevard, Nashville, TN 37208, USA

Abstract

Replication of the human genome relies on the presence of thousands of origins distributed along each of the chromosomes. The activation of these origins occurs in a highly regulated manner to ensure that chromosomes are faithfully duplicated only once during each cell cycle. Failure in this regulation can lead to abnormal cell proliferation, or genomic instability, the hallmarks of cancer cells. The mechanisms determining how, when, and where origins are activated remains a mystery. However, recent technological advances have facilitated the study of DNA replication in a genome-wide scale, and have provided a wealth of information on several features of this process. Here we present an overview of the current progress on our understanding of the initiation step of DNA replication in human cells, and its relationship to abnormal cell proliferation.

Introduction

DNA replication, a vital process in all biological systems, ensures the faithful transfer of genetic information from parent to progeny organisms, and constitutes the major determinant of cell proliferation. This process can be arbitrarily divided into three steps: Initiation, which activates the template DNA to facilitate the incorporation of the first nucleotide in the daughter strands; elongation, which allows further incorporation of nucleotides; and termination, which culminates with the production of two equal copies of the parental DNA. Of these three steps, initiation is the ultimate and critical step in the regulation of cell proliferation.

The replicon model proposed nearly 50 years ago by Jacob, Brenner, and Cuzin [1], has served as a good paradigm for understanding the initiation step of DNA replication. According to this model, initiation of DNA replication depends on the interaction of trans-acting factors (initiators) with cis-acting DNA sequences (replicators or origins). The origin, and adjacent DNA sequences whose replication depend on it, define an independent unit of replication, or replicon. Based on studies on the single replicon present in E.coli, the role of the initiator protein(s) has been expanded to not only to mark the position of the origin, but as a recruitment factor that facilitates the opening of the DNA helix, a step required for the initiation of DNA synthesis [2]. In contrast to bacterial genomes, eukaryotic genomes are composed of thousands of replicons, thus, their replication occurs in a sequential fashion [3-5]. Eukaryotic replication typically follows three cardinal rules: (a) replication is restricted to the S-phase of the cell cycle; (b) replicons are activated in a specific temporal order, some initiating early, and others late in the S phase [6], and (c) for any given replicon, replication occurs only once per S phase and thus re-initiation is prevented [7-8].

Initiation of Eukaryotic DNA Replication

The dissection of the initiation step of DNA replication in eukaryotes requires the answer to three specific questions; How are DNA sites containing origins activated only once during each cell cycle?; What determines When origins fire during the S phase?; and Where are origins localized along each of the chromosomes? At present our knowledge of the how, when, and where of initiation of DNA replication is not complete. A great deal is known about how origins are activated, but less information has been gathered about the underlying mechanisms behind the temporal order of origin firing, as well as, about their chromosomal distribution. Therefore, a coherent picture regarding the relationship between these properties, and how this provides the framework for its regulation, remains to be elucidated. In this paper I wish to present an overview of the current literature in the field with a focus on research in mammalian systems, and in particular, human cells. More comprehensive reviews dealing with several aspects of the initiation process in eukaryotes, methodological approaches used, and the potential misregulation of this process leading to cancer, are available. These include an excellent collection grouped in a special January 2010 issue of Chromosome Research [9], as well as others published elsewhere [10-14].

Activation of origins of DNA replication

Most of our current understanding about how origins are activated relies on the wealth of experimental information obtained studying this step in both the budding yeast S. cerevisiae and Xenopus oocytes. It turns out that in contrast to E.coli, where a single protein (DnaA) is required to load a helicase (DnaB) to make the double helix accessible to the DNA synthesizing machinery; in eukaryotes, this is achieved through a complex process that requires the formation of a multisubunit protein complex at replication origins. As shown in Figure 1, during the M to G1 transition of the cell cycle, a six subunit protein complex named, the origin recognition complex (ORC) binds to origin sequences, and serves as a landing pad for the assembly of other proteins known to be critical for the initiation step in DNA replication including Cdc6, Cd1, and the MCM2-7 complex [15]. This multiprotein-DNA complex, named the pre-replicative complex (pre-RC), allows chromatin to be ’licensed’ for replication in the subsequent S phase. Upon passage to the S phase, preRCs are activated by the action of specific S-phase cyclin-dependent kinases, and Cdc7/ASK kinase. These kinases, promote the recruitment of the GINS complex [82], and Cdc45 to MCM-2-7, triggering the DNA helicase activity of the MCM-2-7 complex [83-84]. The opening of the DNA helix and its stabilization by the binding of RPA facilitates the recruitment of DNA replication enzymes to begin DNA synthesis. The interplay of S phase kinases with other components on the pre-RC also prevents the reformation of this complex, thus ‘licensing’ occurs only once per cell cycle at any given origin [15-16]. In animal cells the inactivation of the licensing system upon entry into the S phase is

*Corresponding author: Manuel S. Valenzuela, Department of Biochemistry and Cancer Biology School of Medicine Meharry Medical College 1005 D.B. Todd Jr. Boulevard Nashville, TN 37208, USA. Tel: 615-327-6253; Fax: 615-327-6442. E-mail: mvalenzuela@mmc.edu

Received January 02, 2012; Accepted February 03, 2012; Published February 08, 2012

Citation: Valenzuela MS (2012) Initiation of DNA Replication in the Human Genome. Hereditary Genetics S1:003. doi:10.4172/2161-1041.S1-003

Copyright: © 2012 Valenzuela MS. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
with some basic chromosomal features such as Giemsa banding, at the chromosomal level, early activation has been associated DNA, and for DNA structures containing forks and replicative bubbles

In agreement with this hypothesis, nuclear matrix preparations from has been proposed to serve as the skeleton for replication factories.

structure which resists high salt extraction (named the nuclear matrix) has been simultaneously activated in a replication factory. While the physical

it has been revealed that components of the origin recognition complex (ORC) recruits Cdc6, which in turn promotes loading of MCM proteins onto chromatin. Upon passage to the S phase, activation of the Cdc7Cdc8 kinase and S phase-promoting cyclin-dependent kinases (CDKs) induces a conformational change in the MCM complex that is required for unwinding of origin DNA by the recruitment of Cdc45 and GINS to the pre-RC. Initiation of DNA synthesis occurs when replication protein A (RPA) and DNA polymerase α-primase, are recruited to the unwound replication origin.

primarily through the regulation of Cdt1. This is achieved in two distinct ways: First, Cdt1 is subject to proteolysis as the cell cycle progresses through S and G2 [17-19]. Second, Cdt1 is inhibited by geminin which specifically binds to Cdt1 during S, G2, and early mitosis [20-22]. The temporal separation of licensing and firing of origins creates a window of single opportunity for origin activation. DNA synthesis does not start until all pre-RCs are formed, and conversely pre-RC formation is inhibited during the DNA synthesis phase of the cell cycle. This provides the cell with another layer of control which prevents both, insufficient, and over-replication. The identification of structural and functional homologs of proteins involved in the initiation of DNA replication in multiple species prompted the speculation that these proteins, or at least ORC components, might bind to a common DNA sequence or motif present in all eukaryotic origins. However, such motifs have yet to be discovered. Nevertheless, it has revealed that components of the pre-RC are positioned downstream of complex networks of extra and intracellular signaling pathways that control cell growth, a finding that has elevated the status of pre-RC components as novel biomarkers of cell growth [23].

Temporal order of origin activation

It has been largely recognized that not all origins are activated at the same time, thus, during the duration of the S phase some origins fire early, others at mid S phase, and others still much later. At the nuclear level, a coordinated activation of origins has been observed, leading to the formation of replication foci, which depending on their timing of activation in the S phase, occupy distinct nuclear territories. It is believed that each focus contains hundreds of origins of replication. While the physical characterization of these putative factories is lacking, a nucleoprotein structure which resists high salt extraction (named the nuclear matrix) has been proposed to serve as the skeleton for replication factories. In agreement with this hypothesis, nuclear matrix preparations from mammalian cells have been found to be enriched for newly synthesized DNA, and for DNA structures containing forks and replicative bubbles [24-26]. At the chromosomal level, early activation has been associated with some basic chromosomal features such as Giemsa banding, G-C content and gene density [27]. In turn, late replication has been associated with chromosomal centromeric and telomeric regions. These findings point once more to the coordinated manner in which large chromosomal domains replicate simultaneously. Studies with individual genes have revealed a close correlation between transcription and early origin activation [28-30]. In general, housekeeping genes, which are constitutively transcribed in most cell types, replicate at early stages of the S phase, whereas the gene expression specificity of tissues determines the timing of replication of each gene. Thus, for a given tissue, expressed genes are replicated early whereas, repressed genes are replicated late, indicating that cellular development plays an important role in the temporal selection of origin activation.

Recent application of genome-wide analysis tools to query the relationship between transcription and replication timing in human cells [31-36], has confirmed that early replication correlates with transcriptional activity. However it has also revealed that this correlation is not air-tight. Thus some regions where active transcription is present appear to replicate late, conversely some silent regions occupy early replicating domains. One interesting example are CpG islands containing genes whose transcription persists even when located in late-replication zones [32-33,35,37]. These findings suggest that there may be other factors other than transcription that regulate the timing of replication.

Origin specification in eukaryotes

The DNA sequences required for an origin specification vary considerably among eukaryotic organisms. At one end of the spectrum, in the baker's yeast S. cerevisiae, three or four 10-15 base pairs spread over a 100-150 bp DNA region are sufficient to function as an origin. These sequences include an essential but degenerate A-T rich11 bp sequence, named autonomously replicating consensus sequence (ACS), and less conserved but accessory DNA sequences called B elements which contain a DNA unwinding element (DUE) that presumably facilitates DNA melting [38]. In the fission yeast S.pombe, the minimum DNA region providing origin function extends to over 800-1000 bp. No ACS element has yet been determined; however several A-T rich sequences of 20-50 bp appear to be important for origin function in S.pombe [39]. Metazoan origins are even less defined and can extend over thousands of base pairs [40]. A requirement for specific DNA sequences for the origin of DNA replication in mammals has been suggested for a few chromosomal sites including a 5.7kb region of the amplified hamster dihydrofolate reductase domain [40-45], an 8kb DNA region in the human β globin gene [46-48], a 2.4kb fragment of the c-myc gene [49], and a 500 bp fragment downstream from the human lamin B2 gene [50-51]. Within each domain however, conflicting evidence has been accumulating suggesting that discrete DNA elements may not be required to initiate DNA replication. Finally, at the other extreme of origin sequence specification, the early embryos of D. melanogaster and X.laevis appear to require little or no sequence specificity to initiate DNA replication, which has been interpreted as a requirement for an extremely rapid S phase in these developmental stages [52]. From all the above, it can be said that although some DNA sites are consistently used as origins of DNA replication, the manner in which these sites are selected, and the sequences that determine their location in eukaryotic cells, remain still elusive. To gain a broader understanding about how DNA sequences are selected for origin specification, DNA microarray and more recently, second generation DNA sequencing technologies have been used to map out the location of origins of DNA replication in the human genome and correlate this location with other chromatin features. We will first describe some of the experimental approaches used to achieve this goal.
Methods to detect replication origins in a genome-wide scale

Two major experimental approaches have been used to identify and localize DNA sequences involved in the initiation step of DNA replication. 1) The first method for identifying sequences containing origins of DNA replication relies on the binding of pre-RC components to DNA to detect origins sequences. The methodology of choice in this case has been chromatin immunoprecipitation (ChIP). Briefly, chromatin is chemically cross-linked in situ and later fragmented into smaller pieces. The resulting chromatin fragments are then immunoprecipitated with appropriate antibodies directed to the pre-RC component of choice. The DNA sequences contained in these fragments are then identified by PCR-based amplification of suspected targets, by hybridization to selected DNA microarrays, or by deep DNA sequencing. 2) The second approach identifies newly synthesized DNA at early stages of replication activation. This approach detects initiation sites for DNA replication which may or not coincide with origin sequences. Unfortunately, both terms have been used interchangeably in the literature. Several methodologies have been used to identify newly synthesized DNA including leading-strand analysis, nascent-strand abundance, density transfer, or gel trapping of replicating DNA [36, 53-59; for further description see refs. 12 & 60]. Of all these methods, the most commonly used is the nascent DNA strand abundance assay which depends on the isolation of short DNA fragments contained in newly activated replications. A variant of this approach as applied to DNA microarray technology is illustrated in Figure 2. Total DNA is collected from an asynchronous population of proliferating cells. Nascent DNA strands are then released by heat denaturation, size fractionated on a 5-30% sucrose gradient to select a pool of fractions containing DNA in the 0.7-1.5 kb size range. This pool is then subjected to digestion with λ-exonuclease in order to digest contaminant DNA fragments that do not harbor an RNA sequence at their 5' ends [61]. This fraction constitutes the test DNA. Total genomic DNA, obtained from the same cell line and sonicated to a similar size range constitutes the reference DNA. Usually both the test and reference DNAs are assayed by real time PCR to confirm the enrichment of bona fide origins relative to adjacent non-origin sequences in the test DNA sample. Both test and reference DNAs are then labeled with Cy-5 and Cy-3 dUTP derivatives, respectively and hybridized to a custom made DNA tiling microarray containing 50-60 nt DNA probes staggered in 50-60 bp steps and spanning selected chromosomal regions. Signals of high Cy-5/Cy-3 ratios are then registered at potential sites of origin activation and their positions assigned to the annotated sequence of the human genome. This methodology provides comprehensive information about initiation sites which are active throughout the S phase, and thus does not require prior synchronization of the cells.

It is important to note however that these approaches are not perfect. Since arrays can only be built to cover regions where hybridization is unique (i.e. regions containing repetitive DNA sequences are excluded), and that this coverage is limited given the size of the genome, the information obtained with array platforms is incomplete and maybe somewhat biased. In contrast, the application of second generation DNA sequencing technologies [53, 55-59]. Three major observations could be drawn from all these studies. First, several hundred origins of DNA replication could be localized in both genic and non-genic chromosomal regions. The distribution of origins was not uniform along the regions queried, since in some instances large DNA tracts appeared to be devoid of active origins. Second, while the cell lines as present in the population, thus differences in the response of individual cells can not be assessed. An alternative approach that circumvents all these problems is the direct visualization of DNA replication at the level of individual cells by the use DNA coming to stretch pulse labeled DNA to quantify both, the presence of initiation sites, as well as, the direction of replication in single DNA fibers [12]. This approach is quite robust to map origin distribution and elongation rates in anonymous DNA segments. By combining this technique with fluorescent in hybridization (FISH) to mark particular chromosomal regions, it is possible to determine the dynamics of the replication program along this region in a cell population. The major limitation however is the relatively short length of the DNA fibers (less than 1 Mb), making the information obtained to be somewhat biased toward fibers showing shorter inter-origin distances.

Genomic distribution of human replication origins

Classical studies using fiber autoradiography have revealed that the number of origins of replication in early embryos is much larger than in differentiated tissues, indicating that the genome contains an excess of origins, compared to the minimum required to replicate the genome [6]. In fact it has been estimated that the number of potential origins is about 4-5-fold times the number required to complete the replication of the genome. All these origins are supposed to be primed as pre-RCs by the end of the G1 phase of the cell cycle thus, paraphrasing the Jesuit dictum “many are called, but few are chosen” it has been proposed that some pre-RCs are activated upon passage to the S and mark the position of active replications, whereas the remaining pre-RCs are located in regions that are passively replicated by adjacent replications [62]. Recent studies using single fiber analysis have indicated a higher level of plasticity in the way origins are chosen. Origins can be classified as constitutive, flexible, or dormant. Constitutive origins fire upon entry into the S-phase in all cell types; flexible origins can be arbitrarily used in different cells of the same tissue; and dormant origins are those used only under special circumstances [10]. It is believed that the major determinant of origin usage is the completion of a single copy of the genome. Therefore the availability of extra origins offers the genome with enough potential resources to achieve this end, under a variety of physiological conditions.

What constitutes an origin, and what determines its localization? From the way origins are activated via the formation of the pre-RC, the simplest definition of an origin would be as that DNA sequence to which ORC binds. Therefore it was thought that the reconstitution of the purified human ORC complex on DNA could resolve this issue. However, when this task was achieved by Vashree et al [63], the reconstituted human ORC, while functionally active in an X.laevis DNA replication assay, and capable of DNA binding affinity, could not distinguish bona fide origin from non-origin DNA sequences. In fact the highest binding affinity of ORC was found with a synthetic polynucleotide containing A-T rich tracts (a 45-fold increase over origin sequences). These results suggested that origin recognition by ORC may either be DNA sequence independent or that other accessory elements provided DNA binding specificity.

Recent reports have addressed the question of localization of active origins in a large scale by applying microarray or second generation DNA sequencing technologies [53, 55-59]. Three major observations could be drawn from all these studies. First, several hundred origins of DNA replication could be localized in both genic and non-genic chromosomal regions. The distribution of origins was not uniform along the regions queried, since in some instances large DNA tracts appeared to be devoid of active origins. Second, while the cell lines as
Figure 2: Experimental strategy to determine the enrichment of short nascent DNA using DNA microarrays. (A) Hypothetical chromosome region undergoing DNA replication through points A, B, and C will produce nascent strand DNA of increasing size. The shortest DNA, which includes point B, will contain an initiation site for DNA replication. (B) Total DNA isolated from synchronously growing cells is isolated and the nascent strand DNA released by heat denaturation is size fractionated through a 5-30% sucrose gradient. Fractions containing DNA in the 0.7-1.5kb size range are pooled and treated with λ exonuclease. A similar size range DNA is produced by sonication of total DNA. Both DNAs are differentially labeled and hybridized to a custom made high density tiling DNA array. By calculating the hybridization ratios at each one of the probes contained in the array a nascent DNA enrichment peak can be positioned within the genomic regions represented in the array.
Reduced expression of some pre-RC components has also been shown to cause severe chromosomal instability ultimately leading to cancer but only under genotoxic conditions [71-72]. An increased incidence of chromosomal breaks, and chromosome instability was observed in mice harboring a hypomorphic mutation in Mcm4 (Chao3 mutant) upon exposure to a replication inhibitor, and about 80% of Chao3 females developed mammary adenocarcinomas [71]. A similar mutation in Mcm2 in mice also caused genomic instability, but in this case mice developed T- and B-lymphomas [72]. Overexpression of pre-RC components could alter the regulation upon passage of the cells into S phase leading to re-replication, a likely determinant of tumorgenesis. A direct demonstration as to how this deregulation could become oncogenic was obtained through the studies of Gonzales et al. [73], who showed that overexpression of Cdc6 in mammalian cells caused specific hypermethylation of the tumor suppressor INK4/ARF locus. This was achieved by the binding of Cdc6 at an origin located around a regulatory element of the INK/ARF locus and the subsequent recruitment of histone deacetylases which cause heterochromatization around this locus. As a result, repression of p14ARF, p16INK4A, and p14NK4B expression ensued. The product of these genes are important activators of both p53 and Rb tumor suppressor pathways [73]. Thus enhanced cell proliferation is the result of derepression of cell cycle progression coupled to overexpression of Cdc6. To gain a better understanding about the oncogenic potential of pre-RC components, it would be interesting to find out how pre-RCs become deregulated in the first place.

The second area of misregulation leading to cancer could occur by the alteration of the temporal order of origin activation. It is conceivable that the change in the replication of an early or late origin could cause a disturbance in the temporal expression of crucial genes. Alternatively, origins may not function at their required time, leaving tracts of unreplicated DNA (a potential source of chromosomal abnormalities), upon passage through mitosis. Unexpectedly, the application of single DNA fiber analysis to study of the replication program around FRA3B, the most active common fragile site found in human lymphoblasts, has led to the discovery that the strength of this site depends on the proper functioning of late origins of DNA replication present at this region. Lymphoblastoid cells that had been subjected to replicative stress few origins fired at the location of the fragile site. In contrast in fibroblastoid cells, where the fragility at this site is not observed, the site contained several active origins [74]. Thus a tissue specificity on the stability of fragile sites was uncovered, which may depend on a differential activation of late origins of DNA replication occupying this region. It is not clear though how this specificity is determined, but this finding clearly illustrates a connection between the control of replication timing and chromosomal instability.

Finally, there is the possibility that in cancer cells alternative replication programs may be operative. According to this hypothesis, the changes in the distribution and activity of replication origins could constitute important determinants of the abnormal physiology observed in cancer cells. Since the pioneering studies by J.H Taylor [75,76], in the 1970s, it has been recognized that the number of active origins used in the genome could be increased under conditions of limited cell growth. More recently, a more detailed analysis using single fiber DNA technology at a specific chromosomal location has confirmed Taylor's results by demonstrating that under conditions of reduced nucleotide pool levels more origins are active [77]. This compensatory effect...
ensures duplication of the genome while maintaining the duration of the S phase. Several reports have suggested that in both transformed and tumor derived cells the number and/or activity of origins per unit length of DNA is increased [78-80; 85-87]. Unfortunately, many of these studies have been limited in scope, thus support for this hypothesis is still lacking. More recently a real time PCR–based nascent DNA abundance assay was used to compare the location and number of active origins around a 78 kb region of the human chromosome 2q34 in both cancer and normal breast cell lines. The replication profile in three different breast cancer cell lines was found to be significantly different that a normal breast cell line. More origins appeared to be activated in the cancer cell lines, but their distribution was broader compared to normal cells [81]. These results suggest that at least in this chromosomal region more origins are recruited in breast cancer cell lines. The significance of this finding will require a large scale comparison of the replication profile between normal and their corresponding cancer cell homologs. This information, coupled with single DNA fiber studies could provide us with enough information to validate or discard this hypothesis.

Summary and Perspectives

DNA replication of human chromosomes occurs by the activity of thousands of units of replication, or replicons. Replicons function following a temporal program which is developmentally regulated and tissue specific. Each replicon harbors an origin of replication which has been primed by the formation a multi-protein complex (pre-RC) prior to S-phase. There is however more origins present in each chromosome, most of which remain dormant. This origin excess allows the cells to respond to situations of replicative stress by the generation of active replicons from dormant origins. While a great deal has been learned about how origins are primed, less information is available about what determines the temporal order, or the positioning of active origins. The recent application of genome-wide approaches to study these questions, have provided a wealth of information about the relationship of origin activation to transcription and chromatin structure but we still lack a coherent understanding about this process. Given the importance of the initiation step in the regulation of cell proliferation, it is not surprising that some components of pre-RC may have an oncogenic function. There is however much to be learned about the signaling pathways that connect mitogen with these components, and how the expression of pre-RC components is regulated. The application of more massive techniques such as second generation DNA sequencing should provide us more information about the cross-talk of DNA replication with other DNA transactions. The next frontier however will be to understand the relationship of chromosome structure and function, as well as how the spatial interactions between chromosomes and their positioning inside the nucleus influence their function.

Acknowledgements

This paper is dedicated to the memory of Susan H. Armor. The author wishes to acknowledge support from the National Institutes of Health (Grant CA138180), and to thank J. Stewart for help with Figure 2.

References


This article was originally published in a special issue, Hereditary Disorders handled by Editor(s). Dr. X. Long Zheng, University of Pennsylvania, USA