1: Introduction

For genetic information to be passed on from one generation of cells to the next the genome has to be replicated with high fidelity. This occurs during the synthesis (S) phase of the cell cycle. DNA replication is a temporally ordered process with different regions of the genome replicating at different times in S phase. The time at which DNA replication is initiated is a highly ordered process. Replication during the early part of S phase has been associated with chromatin conformation, epigenetic and sequence features of the genome and transcriptional activity. Each of these features are considered below. The main aim of the work presented in this thesis is to use genomic arrays to assess replication timing.

Conventional ways of assessing replication timing are laborious and can only assay small regions of the genome. These are described and compared to the use of microarrays for the assay of replication timing. The utilisation of microarrays to assess the replication timing of *Saccharomyces cerevisiae* and *Drosophila melanogaster* is explained. To assay the replication timing of the human genome an array sampling the genome at a 1Mb resolution was used. To examine replication timing of a whole chromosome a genomic clone microarray covering the q arm of chromosome 22 was constructed.

The arrays constructed to assess replication timing can also be used to detect copy number changes. The use of the arrays to identify deletions in 22q due to rearrangement of the immunoglobulin light chain λ and implicated in DiGeorge syndrome was also investigated.

1.1: Chromatin Conformation

1.1.1: Chromatin Condensation

Every human cell contains over two metres of DNA, packaged into a nucleus 5-20 μm in diameter. To achieve this, the DNA is packaged in a highly ordered process (Figure 1.1) This allows condensation of the DNA but still enables replication and

transcription machinery access. Several groups of proteins are involved in the condensation of DNA and its positioning within the interphase nuclei, and the combined structure of DNA and protein is termed chromatin.

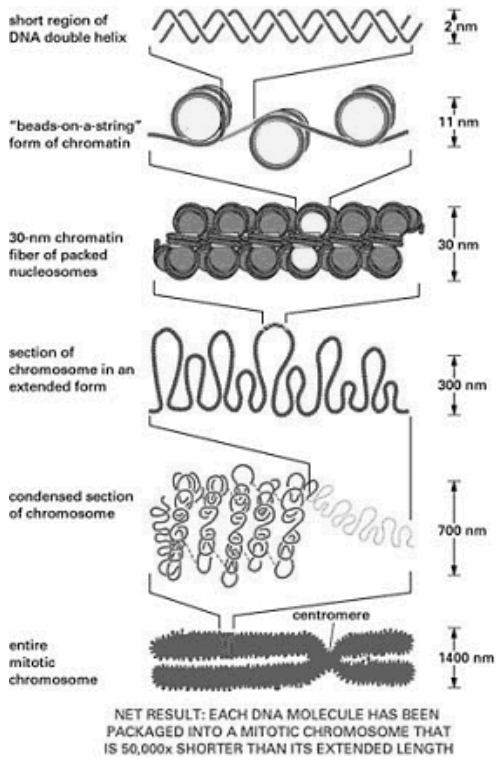

Figure 1.1: Levels of DNA condensation within a eukaryotic chromosome (Strachan 2001)

The first level of DNA condensation occurs when 146bp of DNA is wrapped around a nucleosome. The nucleosome is made of core histone proteins which are described in section 1.1.2. DNA is wrapped twice around a nucleosome to form the recognised beads-on-a-string level of chromatin packing. This produces a fibre approximately 11nm in width.

The second level of chromatin condensation is the arrangement of the nucleosomes into a 30nm chromatin fibre within which there are six nucleosomes per turn. The circular nucleosomes with their DNA wrapped around them can be represented as disc shaped and the discs align so their flat face is roughly parallel to the long axis of the chromatin fibre (See Figure 1.1).

The chromatin fibre is then arranged in loop domains each approximately 120Kb long (Munkel, Eils et al. 1999). These were first seen in scanning electron micrographs of lampbrush chromosomes of the *Urodela* (amphibian) oocyte (Miller 1965). These chromatin loops form rosettes. Each rosette is termed a multi loop sub-compartment (MLS) and comprises of six loops. A further 120Kb of DNA links each rosette. The rosette is attached at its centre to the nuclear matrix (Paulson and Laemmli 1977) and is associated with Histone H1 (Munkel, Eils et al. 1999). Two types of attachment regions have been described; attachment to the nuclear matrix of permanent regulatory regions containing non-transcribed DNA, and attachment to the nuclear skeleton of transient regions of DNA containing transcribing and replicating DNA (Craig, Boyle et al. 1997). This second class of attachment has been visualised by electron microscopy and shown to contain DNA replication and RNA transcription machinery (Hozak, Hassan et al. 1993; Hozak, Jackson et al. 1994).

The final level of chromosome compaction occurs prior to cell division in mitosis. Mitotic chromosomes have a diameter of 700nm. The rosettes are lined up and attached to a central chromosome scaffold of non-histone acidic proteins. The condensation of the metaphase chromosome is due, in part, to a protein called 13S condensin, which is thought to act in an ATP dependent manner and condense chromosomes by inducing a globally positive supercoil. (Kimura and Hirano 1997; Kimura, Rybenkov et al. 1999). The chromosomal scaffold attachment regions are very AT rich and are consequently twice as abundant in the gene poor regions of the chromosome (Craig, Boyle et al. 1997). This tight binding of the AT rich regions to the chromosomal scaffold may explain the banding pattern seen in giemsa stained metaphase chromosomes, with the tightly bound gene poor DNA producing densely stained regions. Metaphase chromosomal bands can therefore be classified into G light (GC rich) or G dark (AT rich) bands.

Chromatin has been classified into two categories, heterochromatin, identified by very intense giemsa staining and euchromatin. In heterochromatin, DNA is in a highly condensed state that restricts the access of additional proteins to the DNA. Heterochromatin was a term first used by the botanist Emil Heitz who identified parts of a moss karyotype that were more compact than others regions (Heitz 1928, Redi, Garagna et al. 2001). These ‘C’ bands were identified by a boiling technique called

'heitzen'. The compact regions were later identified in both animals and plants (Heitz 1930). Unlike its sister euchromatin, heterochromatin was identified as being transcriptionally inert (Ohno 1985). Heterochromatin is usually AT rich, late replicating and gene poor. The condensed state means access to the DNA by other proteins is restricted. Heterochromatin is very rich in repeat sequences such as satellite DNA sequences which are required for correct sister chromatid adhesion and chromatin separation during mitosis. Euchromatin is more loosely coiled than heterochromatin. It is gene rich (and transcriptionally active), and has a high GC content (See Table 1.1). Cimbora *et al*. (Cimbora, Schubeler et al. 2000) found that at the β-globin locus, which is located in euchromatin, the open chromatin state was necessary for early replication.

1.1.2: The Nucleosome and Epigenetic regulation.

Histones are the protein subunits that make up the nucleosomes. They are found in the chromatin of all eukaryotic cells and are highly conserved throughout evolution (Li 2002). The core histones that make up the nucleosome are H2A, H2B, H3 and H4. Two copies of each can be found within each nucleosome and are assembled as illustrated in Figure 1.2. A further class of Histone (H1) can be found associated with the DNA linking the individual nucleosomes.

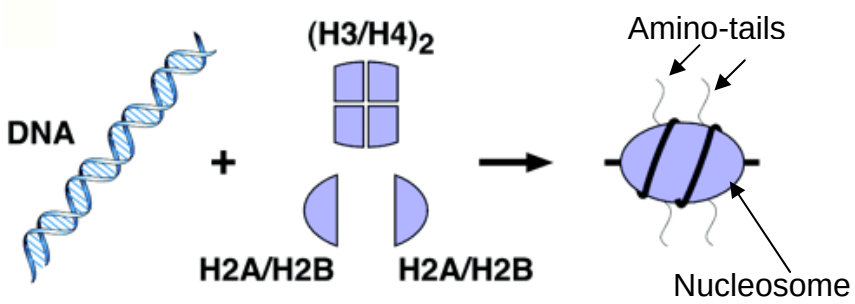

Figure 1.2: Composition of a nucleosome, the fundamental unit of chromatin. Adapted from Grewal *et al*, 2003. (Grewal and Moazed 2003). Each nucleosome contains 146bp of DNA wrapped around an octamer of core histone proteins. Histone proteins are arranged to have their amino tails protruding from the nucleosome core.

The extension of the amino terminus tails from the nucleosome core allow epigenetic regulation of the amino acid residues they contain. Covalent modification of the

amino acid residues is performed by chromatin re-modelling and chromatin modification enzymes. These enzymes mainly target the lysine residues of the amino acid tails. Modification is by acetylation, phosphorylation, methylation and ubiquitylation. The modifications are associated with the transcriptional activity of the associated DNA. Epigenetic regulation of the genome is a heritable feature, yet is independent of DNA sequence (Li 2002).

The amino termini of the H3 and H4 subunits are particularly involved in epigenetic regulation. Figure 1.3 shows the most prominent protein modifications that can occur at the lysine residues within the histone tails. The pattern of histone tail modification is called the histone code.

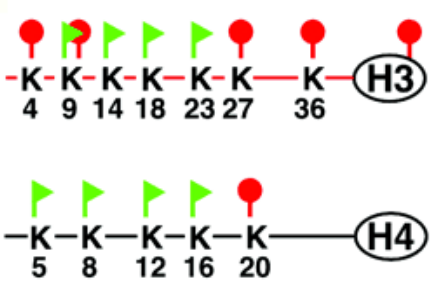

Figure 1.3: Modification at the lysine (K) residues in the H3 and H4 tails reproduced from Grewal *et al* 2003 (Grewal and Moazed 2003). A green flag signifies the lysine is subject to the addition of an acetyl group, whilst the red signal indicates that lysine is subject to the addition of a methyl group.

The addition of acetyl groups to the histone tails is associated with opening of the nucleosomes and a more transcriptionally active state of DNA. The de-compaction of the nucleosomes also makes the DNA more susceptible to DNAase1 activity (Kerem, Goitein et al. 1984). Conversely the removal of acetyl groups from the histone tails results in a closed conformation and transcriptional repression. A correlation between acetylation status and DNA replication timing has also been observed. Hyperacetylated DNA is early replicating, whilst hypoacetylated DNA is late replicating (Vogelauer, Rubbi et al. 2002).

Methylation of the histone tails has different effects, dependent on the location of the lysine residue that is modified. Methylation at the lysine in position 4 of the H3 (H3-K4) tail is associated with active gene expression whilst methylation at H3-K9 tail is involved in heterochromatin assembly. Methylation of H3-K9 is also thought to maintain DNA methylation (Grewal and Moazed 2003).

In *Arabidopsis thaliana*, H3-K9 methylation was shown to promote DNA methylation through heterochromatin protein 1 (HP1) (Jackson, Lindroth et al. 2002). The binding of HP1 to H3-K9 recruits a DNA methyltransferase, which in turn results in a covalent addition of a methyl group to cytosine molecules within the DNA double helix. Although this link is yet to be confirmed in mammals, H3-K9 methylation has been shown to recruit HP1 (Bannister, Zegerman et al. 2001; Lachner, O'Carroll et al. 2001). The down-stream interactions with DNA methylation are likely to be more complex due to the greater variety of methyltransferases in mammalian cells. The methylation of cytosine molecules is particularly important at CpG islands (clusters of the CpG dinucleotide). About 60% of human genes have CpG islands located at their 5' ends (Cross, Clark et al. 2000). Covalent addition of a methyl group to the cytosine of CpG dinucleotides in CpG islands renders their associated genes transcriptionally silent.

Table 1.1: A summary of chromatin and associated epigenetic features.

Feature	**Euchromatin**	**Heterochromatin**
Replication Time	Early	Late
Gene Density	Dense	Sparse
***Alu* Repeat content**	Dense	Sparse
Acetylation of histones	Hyperacetylated	Hypoacetylated
Methylation of CpG	Hypomethylated	Hypermethylated
DNAase sensitivity	Sensitive	Insensitive

In summary, the packaging of DNA into the interphase nucleus is a highly ordered process. This packaging is influenced by the epigenetic modification of the genome by histone acetylation and methylation, and methylation of CpG dinucleotides.

1.2: DNA Replication, the Eukaryotic Cell Cycle and Replication Origins

1.2.1: The Eukaryotic Cell Cycle

About three million cells are replaced and renewed in the human body every minute (http://www.nobel.se/medicine/educational/2001/). This is achieved by cell division. As this takes place the genetic material from the mother cell is divided into two daughter cells. To ensure no genetic material is lost or gained as cell division takes place the entire genome must be copied once and only once before cell division. The doubling of the genome takes place within the synthesis (S) phase of the cell cycle. The cell division and therefore halving of the genetic material takes place in the mitosis (M) phase of the cell cycle. These two phases are interspersed with two growth periods (G1 and G2), to complete the somatic cell cycle (Figure 1.4).

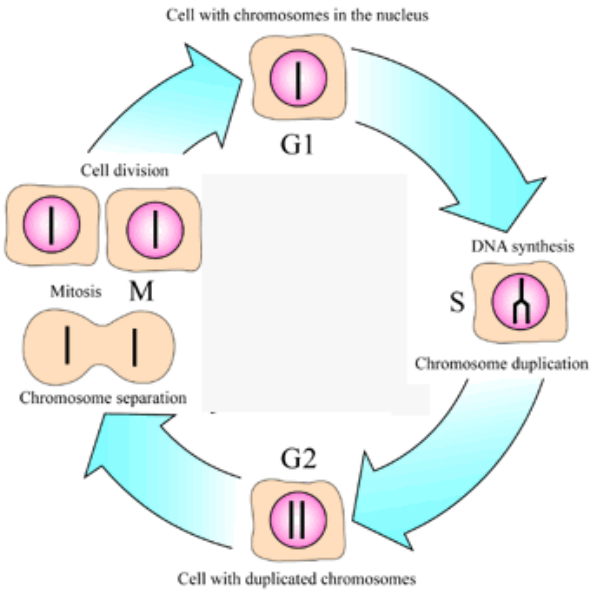

Figure 1.4: The Cell cycle.

1.2.2: DNA replication during the S phase of the cell cycle.

DNA replication is initiated at a specific site called the Replication Origin. As shown in Figure 1.5, a DNA helicase unwinds the DNA to produce a replication fork and DNA is then synthesised in a 5'-3' direction. DNA polymerase δ synthesises the DNA on the leading strand, whilst DNA polymerase ε synthesises short fragments (Okasaki fragments) of DNA using the lagging strand as a template. DNA ligase joins the newly synthesised lagging strand Okasaki fragments together to form a continuous DNA molecule. This continues in a bi-directional fashion until an entire replicon of 40-300Kb has been replicated (Natale, Li et al. 2000). DNA replication is said to be semi-conservative, with each newly synthesised DNA molecule containing one strand from the mother DNA molecule, and one newly synthesised daughter strand.

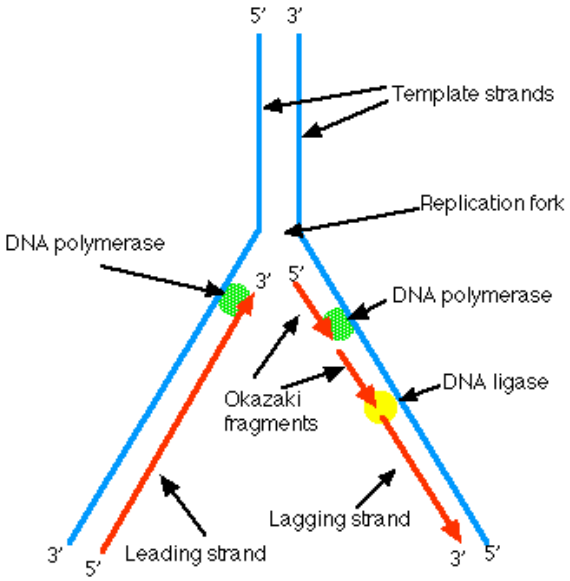

Figure 1.5: The DNA replication fork. For details see text. Figure taken from http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/D/DNAReplication.html

The temporal order of replication is strictly regulated, with some regions of the genome replicating much earlier than others. The replication of one replicon is triggered by the activation of one replication origin. As replication of a replicon nears completion the bubbles of nascent DNA formed from individual origins fuse to form two new DNA molecules (Figure 1.6)

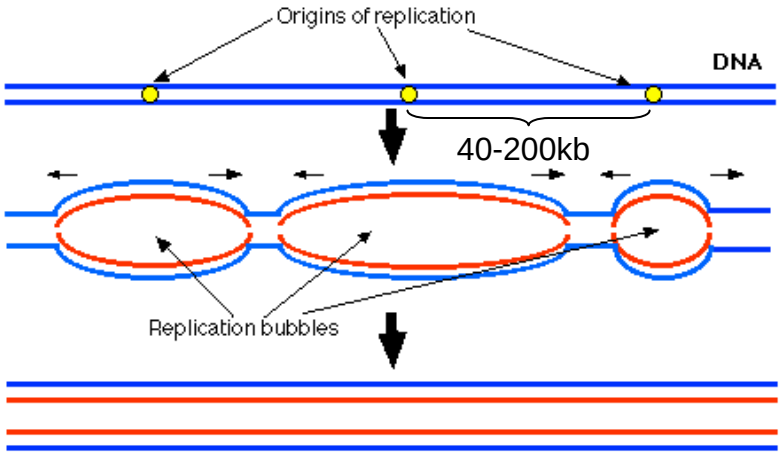

Figure 1.6: DNA replication from origins of replication. Figure taken from http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/D/DNAReplication.html.

In *Escherichia coli*, *Saccharomyces cerevisiae* and *Saccharomyces pombe* replication origins are sequence specific. Although the sequence is not conserved throughout these species it is always an AT rich sequence. However, in higher eukaryotes there is no consensus sequence at which DNA replication is initiated (Kelly and Brown 2000). In higher eukaryotes many sites have the potential to become origins. These sites have specific proteins bound to the chromatin throughout the entire cell cycle, called pre-replication complexes (Pre-RC's). Activation of a few of these sites by the binding of the proteins of an origin recognition complex (ORC) determines the site of replication initiation (Diffley and Labib 2002; Nishitani and Lygerou 2002). Very few replication origins and ORC binding sites have been defined in mammalian cells. Investigation by Natale *et al* showed that each mammalian cell contains 10^4-10^5 molecules of ORC. This suggests DNA replication is initiated once every 60-600Kb (Natale, Li et al. 2000).

The features that determine where the ORC subunits bind initially are still unknown in mammals. In common with the replication origins of lower eukaryotes, many replication origins contain tracts of AT rich sequence (Vashee, Cvetic et al. 2003). However *in vitro* binding studies using purified human ORC showed that ORC did not preferentially bind to these AT tracts (Vashee, Cvetic et al. 2003). In contrast the study of CpG island regions in four mammalian genes showed that clusters of the CpG dinucleotide were at the initiation sites for DNA replication. Short nascent

DNAs synthesised *in vivo*, were found to contain CpG islands suggesting the origins of replication were nested inside the CpG islands (Delgado, Gomez et al. 1998).

The origins that have been described in mammals sub-divide into two categories; loci at which replication is initiated at a discrete chromosomal location and loci at which the origins are found over a much larger zone of initiation (Gilbert 2001). The first category of origins includes the lamin B2 locus and the human β-globin locus. The lamin B2 locus was first mapped by Giacca (Giacca, Zentilin et al. 1994) to a 474bp region corresponding to the non-coding 3' end of the Lamin B2 gene. An evolutionary conserved AT rich region was also observed proximal to this region. The location of the lamin B2 origin was further refined and localised to a single nucleotide (Abdurashidova, Deganuto et al. 2000). The human β-globin origin was the first human origin to be mapped and localises to a defined region, less than 4Kb in size (Kitsberg, Selig et al. 1993) between the adult δ-globin and β-globin genes. Study of this region showed that this locus was used for replication in cells that both expressed and did not express β-globin, and that replication from the origin was bi-directional (Kitsberg, Selig et al. 1993). The β-globin origin was shown to be early replicating and it was proposed to be due to the open chromatin structure of the locus. Models of origin activation at this locus indicated long range control of the origin. Cis acting elements within the open chromatin regulate activation of origins over 50Kb away. (Cimbora, Schubeler et al. 2000).

The second class of replication origin consists of those with a large zone of initiation. These zones can be 10-50Kb regions of DNA within which replication begins from several sites. The Chinese hamster ovary cell DHFR locus is the best defined locus in this group. Two dimensional gel analyses of replication intermediates revealed a 55Kb region of initiation between the DHFR gene and the 2BE2121 gene (gene of unknown function) (Dijkwel and Hamlin 1995). This large region has been shown to contain a minimum of 20 origins, each with different efficiencies of activation (Dijkwel, Wang et al. 2002). Studies at this locus show that when the most active origin is deleted, adjacent origins increase or retain their initiation activity (Kalejta, Li et al. 1998). Other metazoan origins that have a dispersed zone of initiation include the *Drosophila melanogaster* Ori D locus (Ina, Sasaki et al. 2001), the rRNA genes

from human (Little, Platt et al. 1993) and the Chinese hamster Rhodopsin origin (Dijkwel, Mesner et al. 2000).

1.2.3: Cell cycle regulation and control checkpoints

DNA is licensed for replication during the G1 phase of the cell cycle and progression into the G2 phase of the cell cycle is prevented until all DNA has been replicated. This is achieved by a series of regulatory proteins and cell cycle checkpoints.

Early in the G1 phase of the cell cycle the nuclei pass through a 'timing decision point' (TDP). It is during this time that the replication timing program of the cell is established. This occurs before the licensing of the DNA for replication (Gilbert 2002). During the TDP, active DNA sequences are repositioned in clusters in the nucleus to provide a favourable environment for gene transcription. These clusters are related to chromosomal domains and contain a high concentration of replication regulators acting in trans (Dimitrova and Gilbert 2000; Li, Chen et al. 2001). After nuclear repositioning the replication timing program of the nuclei is determined. In the Chinese hamster β-globin locus, the determination of late replication within heterochromatin occurs coincidently with its repositioning at the nuclear periphery (Li, Chen et al. 2001). This provides a link between replication, nuclear position and a favourable environment for transcription (Dimitrova and Gilbert 1999). The importance for nuclear position at the TDP is confirmed by the analysis of sequences situated next to the nuclear periphery (associated with late replication, see section 1.3.3). If these loci are moved away from the nuclear periphery after the TDP they remain late replicating (Gilbert 2002).

To ensure that DNA is replicated once and only once, un-replicated DNA must be licensed to replicate. Only licensed DNA can undergo replication. Licensing occurs after the cells have undergone mitosis in the early G1 phase of the cell cycle. To license an origin for replication pre-replicative complexes (pre-RCs) are formed on the DNA (Nishitani and Lygerou 2002). The pre-RC's contain a six subunit protein, the origin recognition complex (ORC), Cdc6/18, Cdt1 and mini-chromosomal maintenance (MCM) proteins. All these proteins are highly conserved in eukaryotic

cells, highlighting their importance in replication. Some ORC subunits are present on the DNA throughout the cell cycle. The complete ORC is formed and then acts as a 'landing platform' for the other proteins involved in establishing the pre-RC. The Cdc6/188 and Cdt1 load first and recruit the binding of the hexameric MCM complex onto the chromatin (Figure 1.7). About 20 copies of the MCM hexamer are loaded onto the DNA at each Pre-RC (Diffley and Labib 2002). Once all proteins are loaded the origin is licensed for replication.

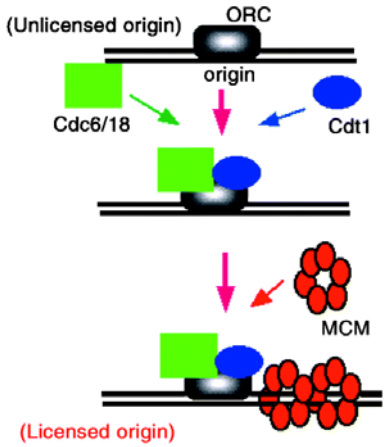

Figure 1.7: Formation of a pre-RC complex to license DNA for replication. Reproduced from Nishitani *et al* (Nishitani and Lygerou 2002).

Before the G1 cells move into S phase of the cell cycle they still have to pass a checkpoint (Restriction point – R) in late G1 phase (Ford and Pardee 1998). Before the restriction point is passed, cells can only continue to cycle if the nucleus receives extracellular signals. Once the restriction point is passed, cells are committed to the S phase of the cell cycle, and ultimately, division (Pollard 2002). This checkpoint is regulated by cyclin dependant kinases (cdk), which are, in turn, dependant on cyclins for their activation.

A number of proteins are involved in the activation of replication although the exact factors determining origin firing are poorly understood (Diffley and Labib 2002). Once an unknown signal to replicate is received, Cyclin D binds with Cdk4/6 and p21 in the cytoplasm to form an active kinase. This can pass through the nuclear envelope and phosphorylate pRB (retinoblastoma protein, a tumour suppressor gene). Passage of pRB through the nuclear envelope leads to progression through the restriction

point. In a second wave of activity, Cyclin E binds to cdk2 which also phosphorylates pRB and leads to progression into S phase (Pollard 2002).

After replication, the cdt1 proteins disassociate from the ORC and are degraded (DePamphilis 2003). This ensures origins that have been replicated no longer have their licence to replicate. The degradation of the licensing proteins also ensures that the cdt1 can not associate with any other ORCs forming a new origin of replication. The MCM and the cdk6/18 proteins also disassociate from newly replicated DNA. The MCMs drop off the DNA in front of the advancing replication fork. The soluble fraction of cdc6/18 is phosphorylated by cyclin dependant kinases and is exported to the cytoplasm after the onset of S phase (Nishitani and Lygerou 2002). The six subunit ORC protein also undergoes some disassociation. Some of the ORC subunits are removed from the chromatin, whilst others stay bound during all stages of the cell cycle. Many other replication control proteins have been identified in yeast and a few of these are conserved in humans. It appears that both cyclin dependant kinases (CDKs) and Skp1-cullin-F-Box protein (SCF) may be involved in re-replication control in human cells (Furstenthal, Swanson et al. 2001) however the exact mechanisms of control are unclear.

There is also a checkpoint control within S phase. Damage to the DNA or stalled replication forks activates a global intra-S integrity checkpoint (Diffley and Labib 2002). Activation of the intra-S checkpoint prevents the firing of origins late in S phase. Entry into mitosis is also prohibited. Experiments producing DNA damage with ionising radiation showed that the protein ATM kinase phosphorylates the proteins Nbs1 and Chk2 to trigger two individual branches of the intra-checkpoint control (Falck, Petrini et al. 2002).

To maintain genome integrity it is also important that no replication occurs in the G2 phase of the cell cycle. The presence of the protein germinin inhibits re-replication of DNA. Germinin binds to Cdt1 tightly; this binding prevents the loading of the MCM protein complex back onto the DNA. Without the loading of MCMs on to the DNA the pre-RC cannot be formed and DNA replication cannot take place (Nishitani and Lygerou 2002).

Finally, in the G2 phase of the cell cycle there is an additional checkpoint. The G2 checkpoint ensures that it is safe for the cell to enter mitotic division and ensures that all DNA replication and repair processes are complete. Rad and Hus (Hydroxyurea sensitive) proteins detect un-replicated DNA. The initiation of this checkpoint control is not fully understood, but a variety of genes are involved. These include Rad9, Rad1 and Hus1, which in human cells form a circular trimer that may detect damage as it slides along the DNA. This ensures only fully replicated undamaged DNA can be incorporated into daughter cells (Pollard 2002).

This evidence about the cell cycle shows that DNA replication is highly regulated. Regulation is applied not only while DNA replication takes place during S phase of the cell cycle, but during the G1 and G2 phases as well. Regulation in the G1 phase licences the DNA for replication and, possibly by organising nuclear position, determines when in S phase DNA replication occurs. Checkpoints in the G2 phase of the cell cycle ensure that all DNA is replicated before cell division takes place.

1.3: Conventional ways of assessing Replication Timing

Replication timing has been measured in a variety of ways since bromodeoxyuridine (BrdU, a thiamine analogue) was first used to assess replication timing by its incorporation into the DNA and assessment in metaphase chromosomes. Methods investigating the banding patterns on metaphase chromosomes have a limited resolution, imposed by chromosome morphology. Other methods, such as quantitative PCR or FISH based methods, assay small regions of the genome at a high resolution but are labour intensive and study of a large region is laborious

1.3.1: Assessment of replication timing by pulse labelling with Bromodeoxyuridine

The replication timing of the whole genome can be assessed at a very low resolution by pulse labelling cycling cells in culture with BrdU (Latt 1973; Yunis 1981; Drouin, Lemieux et al. 1990). The cells are harvested and those containing metaphase chromosomes are examined. The BrdU incorporation can then be used to determine the late and early replicating DNA as BrdU will only be intergrated into the DNA in a

specific fraction of S phase. In the example illustrated in Figure 1.8 the cells were pulse labelled for three hours. The S phase is 8 hours and G2 phase is 2 hours long in the example. In the top example (Figure 1.8), three hours have elapsed between the addition of BrdU to the culture media and the harvest of the metaphase chromosomes. DNA that has replicated in the last hour of S phase can be detected by BrdU incorporation. This is by using a monoclonal antibody to BrdU which is conjugated to a fluorescent reporter molecule. By the variation of the length of incubation with BrdU different banding patterns can be achieved. This relates to the replication timing of the chromosome.

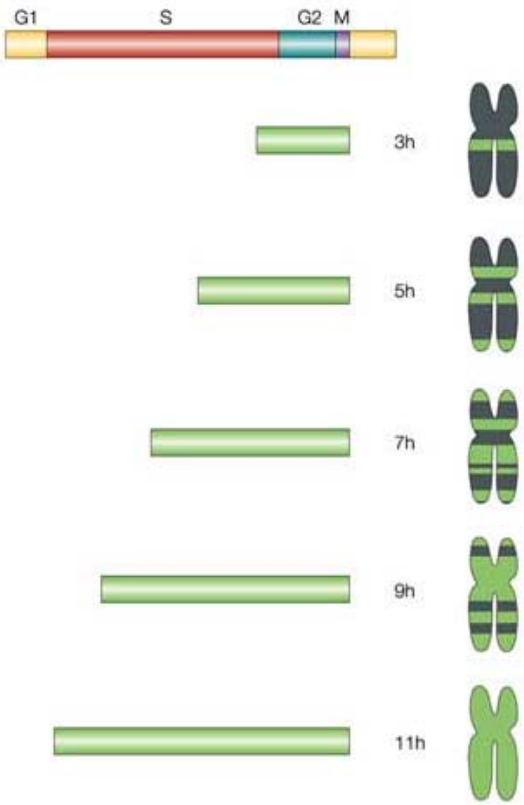

Figure 1.8: Figure from Goren and Cedar 2003. Pulse labelling of a cycling cell line with BrdU leads to banding patterns that report the replication time of the chromosomal bands. Green bands are those regions of DNA that have replicated the most recently before harvest at the end of the S phase. Black regions represent regions of the chromosome that have replicated before the BrdU was added. See text for full details.

The limit to this method is the resolution of the banding pattern of the metaphase chromosomes. Improved banding techniques have enhanced chromosome banding resolution to 1Mb (Yunis 1981; Drouin, Lemieux et al. 1990). However, this still only reports replication timing on a gross cytogenetic level. Using this methodology it is impossible to determine the replication timing of individual genes, or identify subtle changes in replication over small regions.

1.3.2: Assessment of replication timing by fluorescence in situ hybridisation.

An additional method of determining the replication timing of small regions of the genome is to perform a fluorescence *in situ* hybridisation assay (FISH - (Selig, Okumura et al. 1992), reviewed by(Boggs and Chinault 1997)). FISH is a technique that can be used for the mapping of sequences on metaphase chromosomes and the detection of DNA copy number changes within nuclei. DNA from the sequence of interest is labelled incorporation of a hapten which can be detected by conjugation with a flurochrome to make a probe. The DNA probe is hybridised (in the presence of Cot DNA to suppress hybridisation of common repeat sequences) to a DNA target affixed to a glass slide. A fluorescent signal confirms the hybridisation of the probe, and the presence of its complementary sequence within the target.

In the method to assay replication timing described by Selig (Selig, Okumura et al. 1992), an exponentially growing cell line was pulse labelled with BrdU to allow detection of interphase nuclei in the S phase of the cell cycle. The BrdU labelled nuclei were immunodetected with a monoclonal antibody to BrdU conjugated to FITC. The DNA probe was labelled by nick translation with biotin-dUTP and hybridised to the interphase nuclei.

There are three possible hybridisation signal combinations (Figure 1.9). Two singlet signals represent un-replicated DNA at both alleles (Figure 1.9 A), two doublet hybridisation signals represent replicated DNA at both alleles (Figure 1.9 B) and one singlet and one doublet represent one un-replicated and one replicated allele (Figure 1.9 C). This last scenario is usually present in the minority unless the replication

timing at the allele involved is asynchronous. This is further described in section 1.4.4.

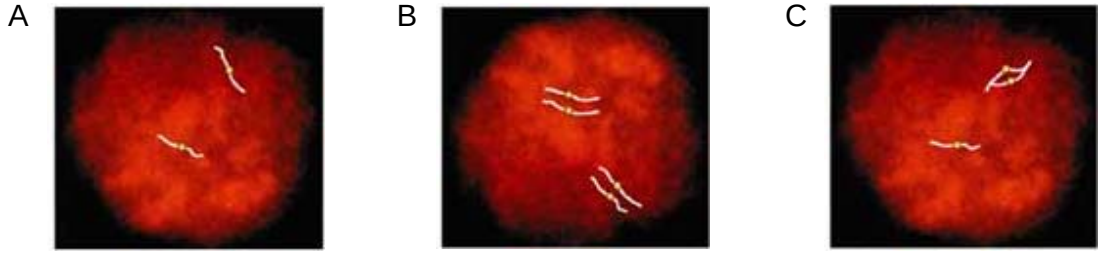

Figure 1.9: Possible DNA replication patterns displayed by the fluorescence hybridisation assay. A: two alleles of un-replicated DNA, B: two alleles of replicated DNA and C: one allele of replicated DNA and one allele of un-replicated DNA. (Adapted from (Goren and Cedar 2003))

The ratio of nuclei with the pattern shown in Figure 1.9B versus the pattern shown in Figure 1.9A indicates the replication time of the loci being investigated. A probe producing a high proportion of doublets reports DNA that replicates early; conversely a probe producing a high proportion of singlets reports DNA that replicates late.

This method can give a wide variation in results, for example, the p53 locus data studied in a normal lymphoid cell line given by Amiel *et al* (Amiel, Litmanovitch et al. 1998) for three different individuals is reported as 47% double singlets (ss), 80%ss and 62%ss. There is also an inherent inaccuracy in the method as the assay relies on being able to resolve a doublet signal from a singlet. Therefore the DNA must separate enough after replication to resolve the doublet signal. Recent evidence suggests that the time between replication and sister chromatid separation is different at different loci (Azuara, Brown et al. 2003). This is due to the selective holding together of sister chromatids by specific protein complexes. Therefore this assay does not measure replication timing as previously thought, but measures sister chromatid separation. This raises the possibility that previous FISH based assays of replication timing may have overestimated the number of late replicating loci.

1.3.3: Replication Timing by flow sorting and PCR

One way to increase resolution is to compare the relative abundance of specific sequences of nascent DNA at different stages of the cell cycle (Gilbert 1986, Hassan and Cook 1993, Sinnett, Flint et al. 1993). Cell cultures in the exponential phase of their growth are pulse labelled with BrdU which is incorporated into the newly synthesised DNA. The nuclei are then stained with propidium iodide and equal numbers of nuclei are sorted into four S phase fractions (Figure 1.10 (Azuara, Brown et al. 2003)). Nascent DNA is then extracted by immunoprecipitation with anti BrdU. The quantity of newly synthesised DNA in each fraction is detected by semi-quantitative PCR with primers specific for the loci of interest. Fractions with the most nascent DNA have the most intense band when the PCR products are run on a gel. This will narrow down the time of replication to one of the four S phase fractions.

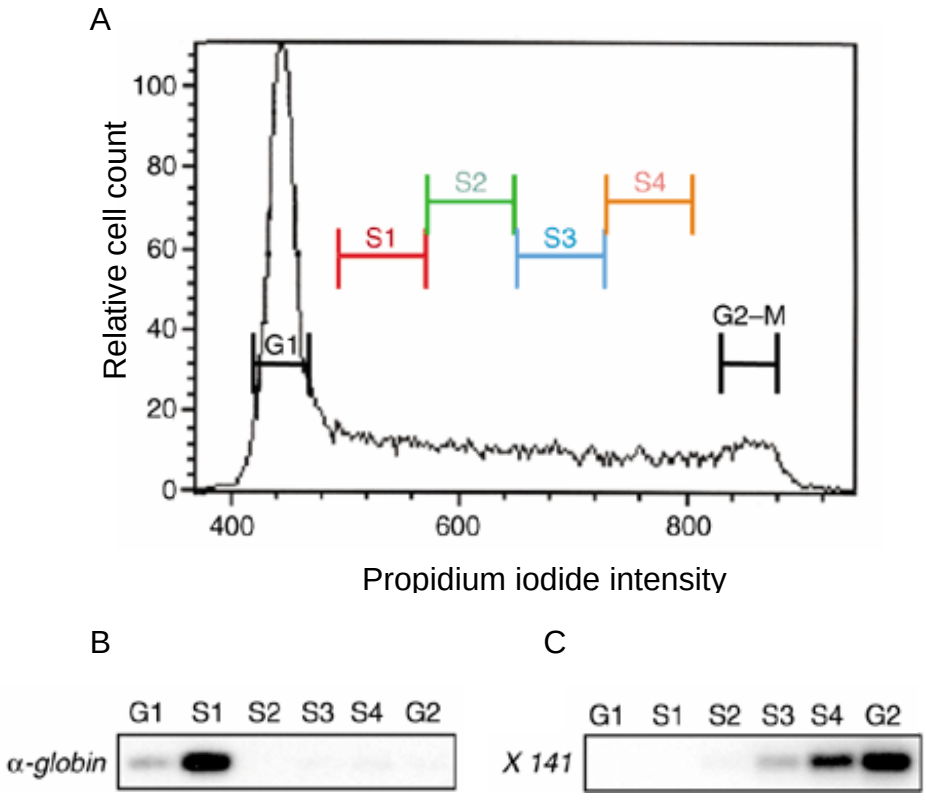

Figure 1.10: (Adapted from Azura et al, 2003) A: Cell cycle profile showing gate positions required to sort the nuclei into four separate S phase fractions. B. Gel photograph illustrating the enrichment in the S1 fraction for an early replicating locus. C. Gel photograph illustrating the enrichment in the S4/G2 fraction for an early replicating locus.

This method is limited by the number of fractions into which S phase can be accurately sorted. It will only therefore, give an approximation of the time of replication, placing the replication time for each locus within a specific quartile of S phase, or between two S phase fractions. This technique is also very labour intensive and is mainly used to screen small regions of the genome. However, recently the technique has been used to screen whole chromosome arms. Wanatabe and co-workers published a replication timing profile for chromosome arms 11q and 21q (Watanabe, Fujiyama et al. 2002). The average resolution sampled in the chromosome 11q data was 300Kb and in the chromosome 21q data this was increased to 200Kb (Figure 1.11). This study shows how the replication timing along the chromosome arm is linked to the GC content and cytogenetic banding. However the method is still limited as S phase was only sorted into four fractions so again the replication time given is an approximation.

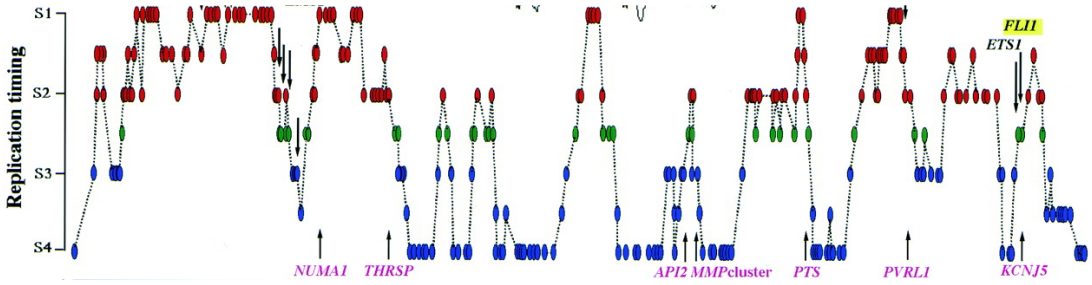

Figure 1.11: Replication timing profile of chromosome 11q using quantitative PCR. Adapted from Figure 3, (Watanabe, Fujiyama et al. 2002).

Ideally, methods for assessing replication timing would sample large regions of the genome at a high resolution and accuracy. To this end, I have investigated the use of genomic arrays to calculate the copy number change associated with DNA replication. (Section 1.6)

1.4: Temporal control of Replication Origin Activation

1.4.1: DNA Replication Timing and Correlation with Sequence Features.

Table 1.1 summarises how genome features such as GC content and gene density are related to the type of chromatin. The sequencing of the human genome (IHGSC 2001) allowed large scale analysis of previously hypothesised links between GC content, gene density, repeat content, cytogenetic banding and recombination rate. Publication of the human genome sequence (IHGSC 2001)of the human genome revealed that 98% of clones mapping to the darkest G-bands have a low GC content (average 37%), whereas more than 80% of clones located in the lightest G-bands are in regions of high GC content (average 45%).

Correlations have also been reported between GC content, cytogenetic banding and replication timing. Pulse labelling replicating cells with Bromodeoxyuridine) and examination of harvested metaphase cells (section 1.3) reveals a banding pattern of replicating DNA. The early replicating, BrdU incorporating, bands correlate with the cytogenetic G-light (GC-rich) bands (Latt 1973; Drouin, Lemieux et al. 1990).

Studies at individual G dark-G light boundaries correlate early replication with a high GC content. A high resolution study assessing replication timing using quantitative PCR analysed the boundary between the G light 13q14.3 and the G dark 13q21.1. The G light side of the boundary was shown to replicate early, whilst the G dark side of the boundary replicated late. Analysis using PCR primers spaced at approx 150Kb intervals showed that the switch in replication timing happened gradually from early-mid-late over a 1-2Mb region, rather than an abrupt change to coincide with the G light-G dark boundary (Strehl, LaSalle et al. 1997).

A further study over the major histocompatibility complex (MHC) region on chromosome 6p21.3 has also shown a change in replication with a progression from a GC poor to GC rich region. Figure 1.12 (Tenzen, Yamagata et al. 1997) shows a 450Kb region across the MHC Class II (GC-poor) and MHC Class III (GC-rich).

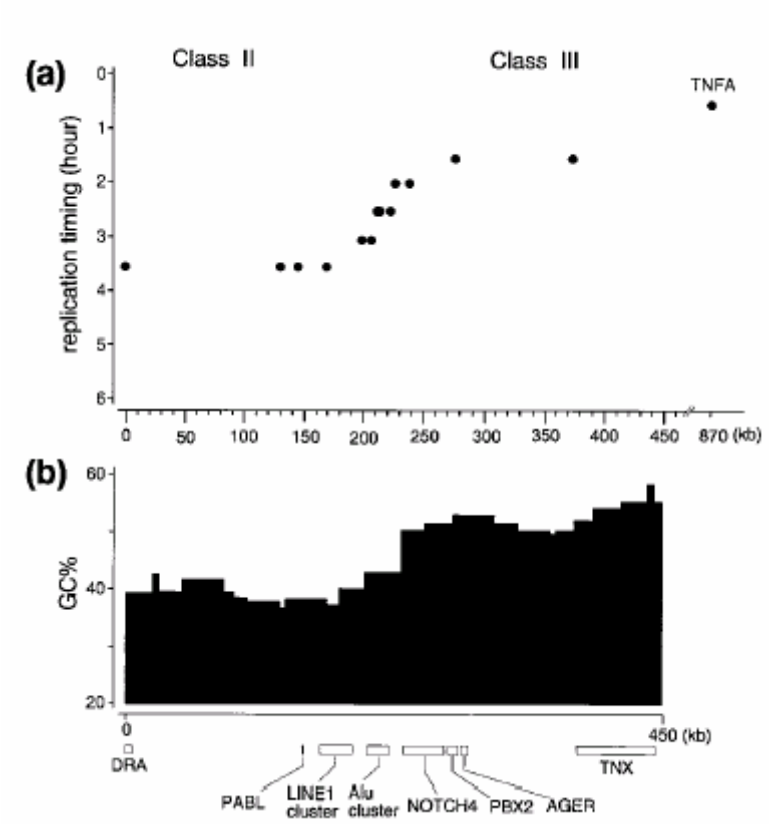

Figure 1.12: Change in Replication Timing (a) and GC content (b) across a 450Kb region of the MHC Class II and Class III taken from (Tenzen, Yamagata et al. 1997).

The Class III region of the MHC replicates about an hour and a half into S phase (after the cells have been removed from an aphedicolin block). The Class II region of the MHC replicates about two hours later (3.5 hours after release from an amphedicolin block). It can be seen that the transition between the change in replication timing and the change in GC content occur at the same point. The transition takes place within a zone of about 100kb, with loci in the transition zone undergoing replication at a mid time point.

The correlation between replication timing and GC content has also been observed over an entire chromosome arm (Watanabe, Fujiyama et al. 2002). Replication timing was assessed on chromosome 11q and chromosome 21q at a resolution of 300Kb (11q) and 200Kb (21q), using flow sorted S phase fractions and PCR (section 1.3). Wanatabe *et al* described a general correlation between replication timing and GC content on both chromosome arms. Zones of early replication were more GC rich than the late zones, although they did observe that the correlation reduced in atypical

regions of the chromosome arms (the pericentric and telomeric regions). The chromosome arm data was also used to study many transition zones between early and late replication (or vice versa). The data showed that the transitions in replication timing were identical, or very close to, regions showing a transition in GC content. The two chromosome arm profiles (Figure 1.13) show the correlation between replication timing and GC content.

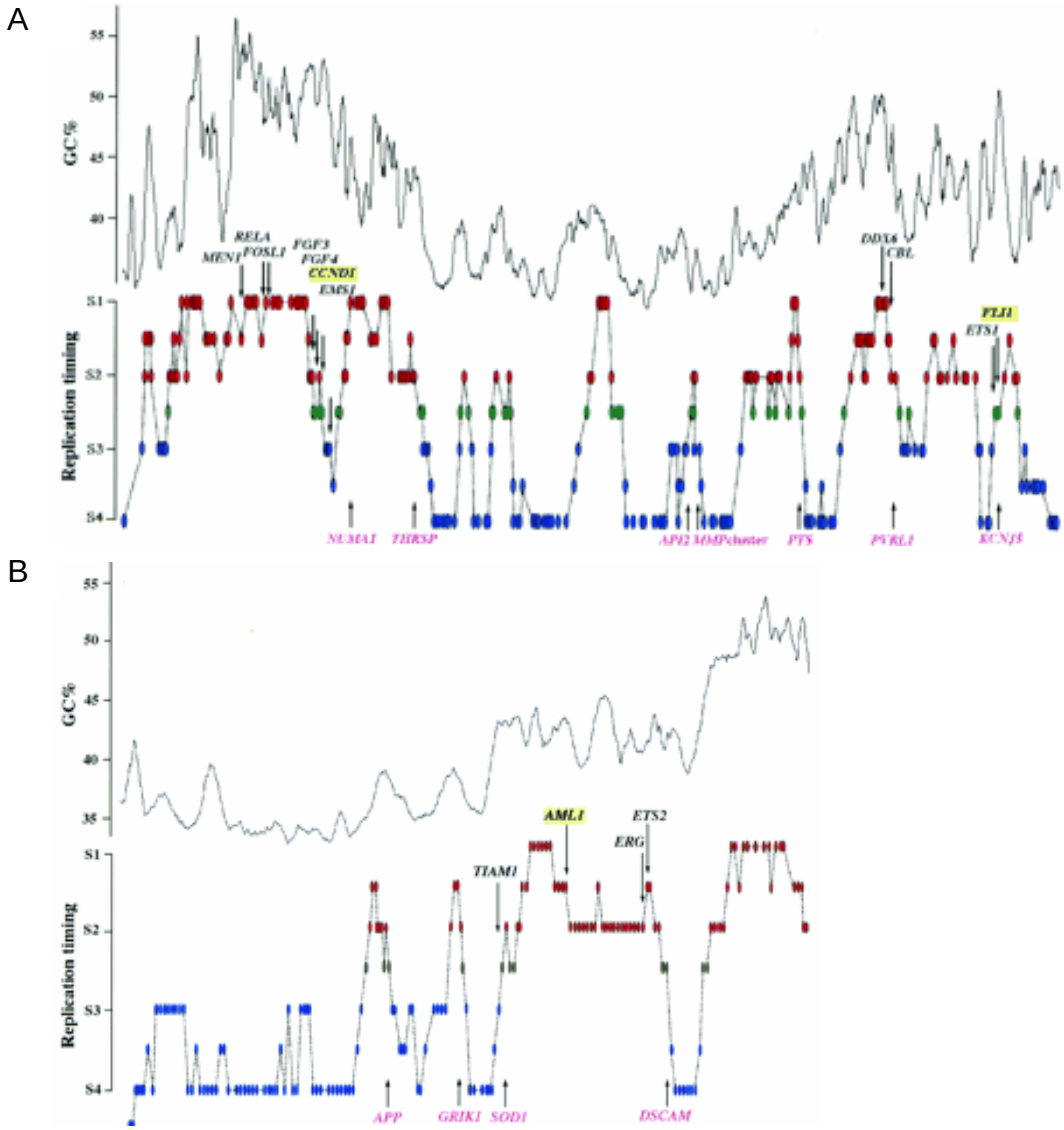

Figure 1.13: The correlation between GC content and replication timing profile on Chromosome 11q (A) and Chromosome 21q (B) (Watanabe, Fujiyama et al. 2002).

The overall correlation between replication timing and GC content can clearly be seen, however it can also be observed that the correlation is not absolute and changes in replication timing can be associated with small local change in GC content.

Many of the loci reporting an atypical relationship between GC content and replication timing are located at the centromeres and telomeres of chromosomes. Heterochromatic centromeric or telomeric DNA contains arrays of repetitive sequence. In *Saccharomyces cerevisiae* all telomeres have been shown to be late replicating. However, genome-wide analysis of the higher eukaryote *Drosophila melanogaster* showed that the euchromatin located close to either the centromere or the telomere was not found to be late replicating (Schubeler, Scalzo et al. 2002). Closer analysis of the centromere on the *Drosophila* chromosome 2L revealed that genes located in the β-heterochromatin of the centromere were early replicating, although they are not transcribed. Study of α-satellite DNA at the centromeres of human chromosomes reveals that although the centromeres replicate at slightly different times in the cell cycle, they all replicate in a narrow window during late S phase (Hultdin, Gronlund et al. 2001).

The investigation of human telomeres located on 22q and 16p13.3 have shown that in common with what was seen in yeast, some human telomeres are late replicating (Smith and Higgs 1999, Ofir, Wong et al. 1999). However, studies of other telomeric regions in the human showed that human telomeric sequences, like those in *Drosophila*, replicate at variable times (Hultdin, Gronlund et al. 2001, Ten Hagen, Gilbert et al. 1990).

Closer analysis of the replication timing of 325Kb of telomeric DNA from 16p13.3 using a FISH based assay (Selig, Okumura et al. 1992) showed the GC rich region lying in the most centromeric region of the 325Kb studied contained widely expressed genes and was early replicating. The subtelomeric 20Kb of the sequence was late replicating. Movement of early replicating DNA adjacent to the heterochromatic telomeric repeats delays the replication of the inserted sequence (Smith and Higgs 1999).

In summary, replication timing is correlated with GC content. The study of repetitive DNA sequence features reveals the heterochromatic telomeric DNA does not seem to

have a specific replication time in human cells. However, examination of the repetitive centromeric DNA shows a clear bias towards late replication.

1.4.2: DNA Replication Timing and Correlation with Chromatin and Epigenetic Features

The investigation of the replication timing in the human β-globin locus correlated replication timing with the open structure of the chromatin. This suggested that chromatin conformation was important in replication timing and that there is long range control of both origin choice and replication timing at the human β-globin locus (Cimbora, Schubeler et al. 2000).

Section 1.1 describes how chromatin condensation in the interphase nuclei is a highly ordered process. Chromatin conformation is associated with a variety of epigenetic features such as acetylation and methylation of histone proteins within the nucleosome and methylation of the CpG dinucleotide at CpG islands. The epigenetic status of the chromatin is reflected in its replication timing. It has long been acknowledged that the tightly condensed, epigenetically silenced, transcriptionally inert heterochromatin is late replicating (Holmquist 1987). In contrast the loosely coiled, transcriptionally active chromatin replicates early. This is particularly evident in the female mammalian X chromosomes and study of alleles that are asynchronously replicated (Discussed further in 1.4.4). The early replicating allele is usually hyperacetylated and the CpG islands are hypomethylated. In contrast the late replicating allele is transcriptionally silent, with methylated CpG islands and hypoacetylated histone proteins. The study of X inactivation in embryonic stem cells reveals that the change in replication timing (to late replication) of the inactive X chromosome is a relatively early event, taking place after the coating of the inactive X with Xist RNA but before changes in histone acetylation, or methylation of CpG islands at promoters occurs (for further details see 1.4.4 – (Avner and Heard 2001)).

The completion of S phase is vital for the complete condensation of the nuclei in M phase of the cell cycle. Replication mutants that are unable to complete S phase have condensation defects (Gatti and Baker 1989). ORC is just one of the proteins that are

important in both DNA replication and DNA condensation. The mitotic chromosomes of *Drosophila* ORC mutants are shorter and thicker than wild type chromosomes (Pflumm and Botchan 2001). Although some levels of chromosome condensation are possible without complete replication it is thought that complete replication is important for lengthwise compaction (Pflumm 2002). This is supported by the longer, less compact mitotic chromosomes observed during embryogenesis. During this time there are many more replication origins present to support rapid cell division. This suggests replication timing may alter in nuclei with chromatin compaction defects. During development, as the loop size between nuclear attachments is increased, the metaphase chromosomes become shorter and thicker (Wang, Castano et al. 2000).

1.4.3: DNA Replication Timing and correlation with Nuclear Position

As discussed previously in section 1.4.1, the correlation between replication timing and DNA sequence indicates there are similar regions of replication timing across the genome. Regions of similar replication timing are found in bands that correlate with the G-bands. There is a further correlation between replication timing and spatial relationship i.e. position within interphase nuclei.

In Figure 1.14, taken from Ferreira *et al* (Ferreira, Paolella et al. 1997). Nuclei were pulse labelled one hour before harvest and at different times after synchronous release into S phase, five different replication patterns are shown. Five different replication patterns can be seen (A-E). A and B show very early replicating DNA in foci dispersed throughout the internal interphase environment stained white on these micrographs. Little replication occurs towards the periphery of the interphase nuclei or adjacent to the nucleolus (The position of the nucleolus can be seen in the corresponding DIC images – F-J). As the interphase nuclei progress through S phase the pattern of replication changes. Figure 1.14 C shows some replication still occurs in internal foci, but most is localised adjacent to the nuclear membranes and the nucleolus. Finally in stages D and E it can be seen that no replication occurs within the internal nuclear environment, with all replication occurring adjacent to the nuclear membrane.

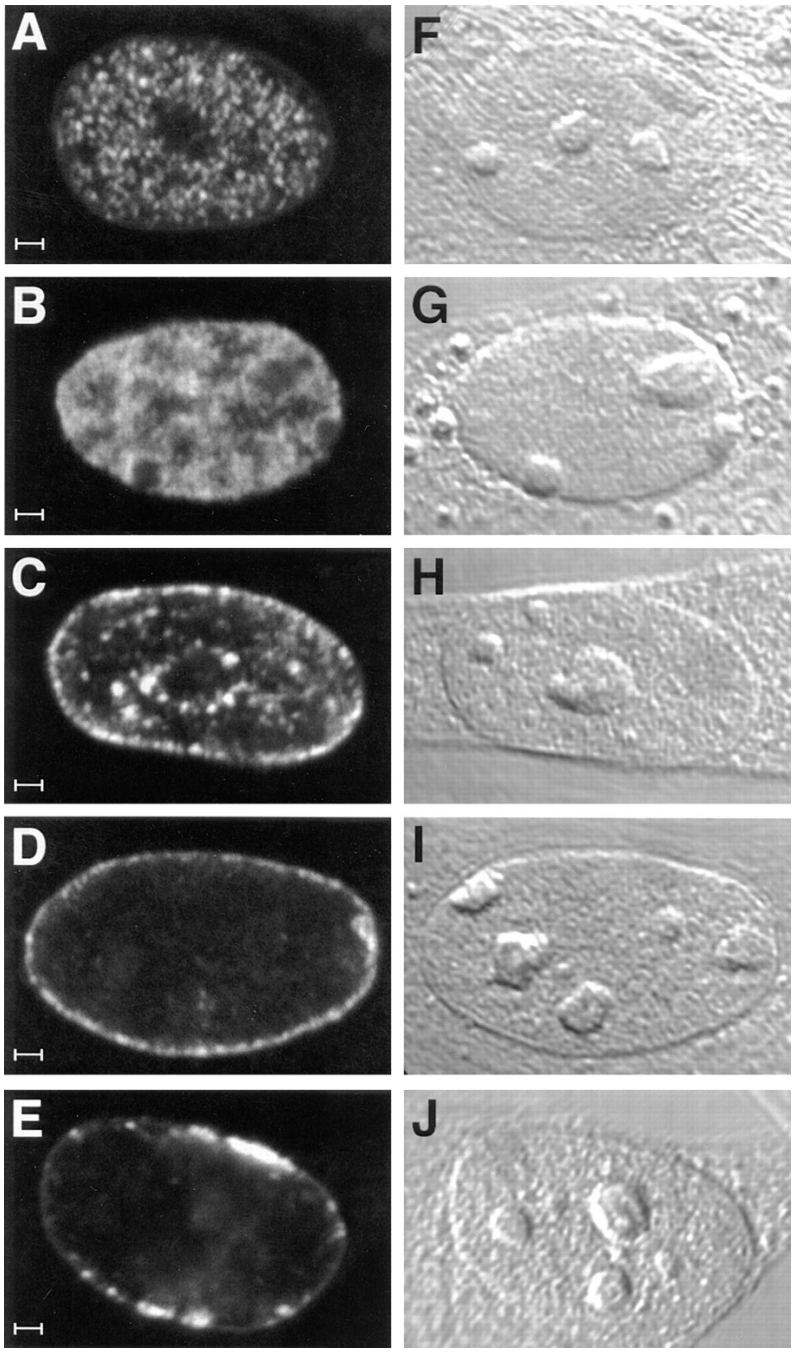

Figure 1.14: Different stages of replication in the interphase nucleus. Figure taken from (Ferreira, Paolella et al. 1997). For details see text.

The localisation of early replicating DNA in the internal nuclear environment and late replicating DNA to the periphery of interphase nuclei has also been visualised by differential pulse labelling of early and late replicating DNA (Schermelleh, Solovei et al. 2001). In Figure 1.15, early replicating DNA has been labelled in blue, whilst mid-late replicating DNA was labelled in red by the incorporation of differentially labelled nucleotides. Again it can be seen that early replication occurs in foci in the inside of

the interphase nuclei, whilst late replication occurs adjacent to the nucleolus and nuclear envelope. Each replication focus contains approximately 0.25 - 1.5 Mb of DNA and the replication machinery required. Each focus takes approximately one hour to replicate (Cremer and Cremer 2001). The foci that are late replicating and localise to the nuclei periphery have also been confirmed as containing AT-rich DNA located in G dark bands (Zink, Bornfleth et al. 1999). Early and late replicating DNA occupy distinct foci within the interphase nuclei. The median overlap between late and early DNA location being only 5-10% (Zink, Bornfleth et al. 1999).

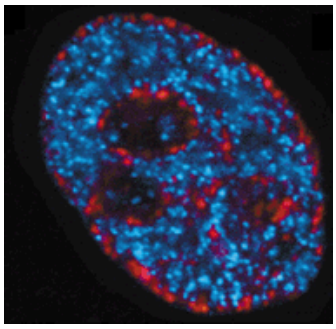

Figure 1.15: Localisation of early replicating DNA (blue) and late replicating DNA (red) in an interphase nuclei (Schermelleh, Solovei et al. 2001).

Similar labelling experiments were carried out on nuclei that were also stained for the protein lamin B. Lamin B is a component of the nuclear lamina and localises to the nucleoplasmic side of the nuclear envelope. It can be seen that late replicating DNA co-localises with lamin B at the nuclear envelope (Figure 1.16)

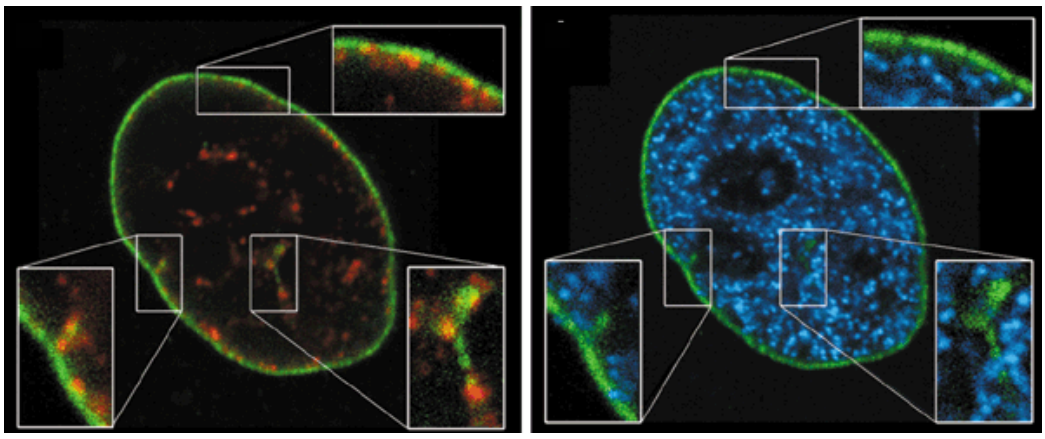

Figure 1.16: Mid-late replicating chromosome domains (red) associate with lamin B (green). Early replicating DNA (blue) does not (Schermelleh, Solovei et al. 2001).

The position of the chromatin within the interphase nuclei is established during early G1 phase of the cell cycle (Gilbert 2001). Studies in Chinese hamster ovary cells show that DNA attached to the nuclear matrix during the G1 phase of the cell cycle is enriched in replication origins (Djeliova, Russev et al. 2001). DNA replication occurs on the nuclear matrix and after replication has occurred the DNA disassociates. Replication origins are therefore transiently attached to the nuclear matrix, associating with replication origins in G1 and disassociating during the S phase. The establishment of the interphase position is co-incidental with the establishment of DNA replication timing as described in 1.2.3.

The position of the immunoglobulin heavy chain locus (IgH) in B cells shows that localisation in the interphase nuclei is dependant on replication timing and gene activity. During early stages of B cell development, the IgH locus is early replicating in both alleles and is maintained away from the nuclear periphery in the centre of the interphase nuclei. Once in the centre of the interphase nuclei VDJ recombination and germ line transcription of the IgH locus occur. In the later stages of the B cell development germ line transcription of the IgH locus is turned off. The entire locus no longer replicates early and the IgH locus is localised to the periphery of the interphase nuclei. The peripheral position of the IgH locus may ensure that the DNA can only replicate at the end of S phase (Zhou, Ermakova et al. 2002). Current data suggests that perinuclear position is indicative but not sufficient for late replication.

In summary, there is clearly an association between the position of DNA within the G1 and S phase nuclei and the time of replication within S phase. The position of the early replicating DNA away from the periphery of interphase nuclei is important. It is also significant that early replicating DNA is attached to the DNA matrix. This ensures the early replicating DNA is in a position favourable to replication when the cell enters S phase.

1.4.4: Asynchronous DNA Replication.

Loci at which expression occurs from both alleles replicate synchronously. Each allele on the two sister chromosomes replicate at the same point in the S phase of the cell cycle. This is not true for loci where expression is monoallelic, as these replicate asynchronously (Mostoslavsky, Singh et al. 2001), (Goren and Cedar 2003). Examples include imprinted regions, the X Chromosomes in females, immunoreceptor genes and genes encoding olfactory receptors (Singh, Ebrahimi et al. 2003).

Early evidence for the asynchronous replication timing in monoallelically expressed regions came from a study of the imprinted Prader-Willi syndrome critical region on 15q11.2 (Izumikawa, Naritomi et al. 1991). Replication banding studies (such as those described in 1.3.1) showed replication asynchrony between homologues of 15q11.2 in about 40% of individuals (Izumikawa, Naritomi et al. 1991).

The asynchronous replication timing of imprinted chromosomal regions seemed to be confirmed at a higher resolution by the characterisation of this region using the FISH assay described in 1.3.2 (Selig, Okumura et al. 1992). Imprinted regions showed a much higher proportion of nuclei displaying the ‘one singlet one doublet’ hybridisation signal (Figure 1. 9c (Goren and Cedar 2003), (Kitsberg, Selig et al. 1993). However, due to the limitations of this assay described by Azuara and colleagues (Azuara, Brown et al. 2003) the assessment of replication timing by the FISH assay may only show the difference between sister chromatid separation and not a difference in replication timing.

Nevertheless, replication asynchrony could be confirmed at a higher resolution using a different method at the human Igf2 loci, located on the imprinted region in chromosome 11p15 (Simon, Tenzen et al. 1999). Replication timing was assessed using quantitative PCR on flow sorted S phase fractions. Restriction site polymorphisms were used to distinguish the maternal and paternal alleles and one chromosomal copy was seen to replicate before the other.

Asynchronous replication is also witnessed in the female X chromosome in mammals. Females have two X chromosomes whereas males only have one X chromosome. To avoid any X chromosome gene dosage imbalance, one of the female X chromosomes is modified in the late blastocyst to become inactive. This inactivation involves chromosome-wide epigenetic changes, making the DNA chosen for inactivation transcriptionally inert. There is also a shift to a later replication time for the inactive X chromosome, whilst the active X chromosome retains its original replication time. This is one of the first developmental changes involved in X inactivation and precedes histone hypoactylation and DNA methylation. About 15% of genes on the X chromosome escape inactivation. It has been suggested that LINE repeats are involved in the propagation of X inactivation along the chromosome. Regions with a lower density of LINE repeats than the rest of the chromosome escape inactivation. (Avner and Heard 2001).

A third category of monoallelically expressed genes are found on autosomes, but unlike imprinted genes the pattern of expression is independent of the parent of origin. These genes include members of the family of olfactory receptors, or encode immuno-receptor genes. In both these systems, a wide range of receptors are encoded in the genome but it is important that only one is expressed in each individual cell (In olfactory receptors this is important for sensitivity to different aromas and in immunoreceptor genes this is important for the clonal development of B cells). An olfactory receptor neuron contains genes for more than 1,000 receptors found within clusters throughout the genome; however, only one is expressed on the cell's surface. To achieve this, the clusters not being expressed are epigenetically silenced, and only one parental allele is expressed in each cell. Part of this epigenetic silencing is reflected in a transition to late replication (Chess, Simon et al. 1994, Singh, Ebrahimi et al. 2003). The genes encoding antigen receptors are also monoallelically expressed. The late replication is randomly established within the early embryo and is maintained by the clonal development of these cells. In B cells it is predominantly the early replicating allele that undergoes rearrangement and ultimately, expression (Mostoslavsky, Singh et al. 2001).

In all three examples of monoallelic expression almost all cases show that the allele that replicates early is the allele that is expressed. This provides a possible link between replication timing and gene expression.

1.5: DNA Replication Timing and Correlation with Transcription

In higher eukaryotes DNA replication timing is thought to correlate with transcriptional activity. Studies of regions of the genome suggest that early replicating DNA is transcriptionally active. Conversely late replicating DNA is transcriptionally inert (Holmquist 1987). Experiments on HeLa cells have compared the patterns of transcription and replication (Hassan, Errington et al. 1994), (Hassan and Cook 1994). Optical sections were taken through the cell cycle of HeLa cells at different stages of the cell cycle. Transcription was indicated by labelling with Texas Red. Replication is indicated by labelling with fluorescein (green) as shown in Figure 1.17. During mitosis the cell is black as no transcription or replication occurs. At sites where replication and transcription occur together the signal ranges from purple to white, as intensities increase. (Figure 1.17) Transcriptionally active DNA was found to replicate in early S phase.

The few ubiquitously expressed housekeeping genes that have been assayed for replication timing have all been located within early replicating G light bands of chromosomes. On the other hand, tissue specific genes are located within G dark bands and almost always replicate late except when the tissue specific region of the genome is expressed; the gene then becomes early replicating. An example of this is the β-globin gene. This gene cluster is in a 200-300Kb stretch of hypoacteylated, DNase-1 resistant late replicating chromatin; however in erythroblasts, which can be induced to express β-globin, the 1Mb of chromatin surrounding the β-globin gene becomes early replicating (Gilbert 2002), (Cimbora, Schubeler et al. 2000).

This phenomenon is particularly evident in developmentally significant genes. Of those that have been studied, at the stage in development in which the gene is expressed it also replicates early. However, once the gene is no longer expressed the replication timing of the gene changes and is delayed (Nothias, Miranda et al. 1996).

It has been proposed that in higher eukaryotes replication timing is a developmentally regulated process that is closely associated with gene expression (Holmquist 1987).

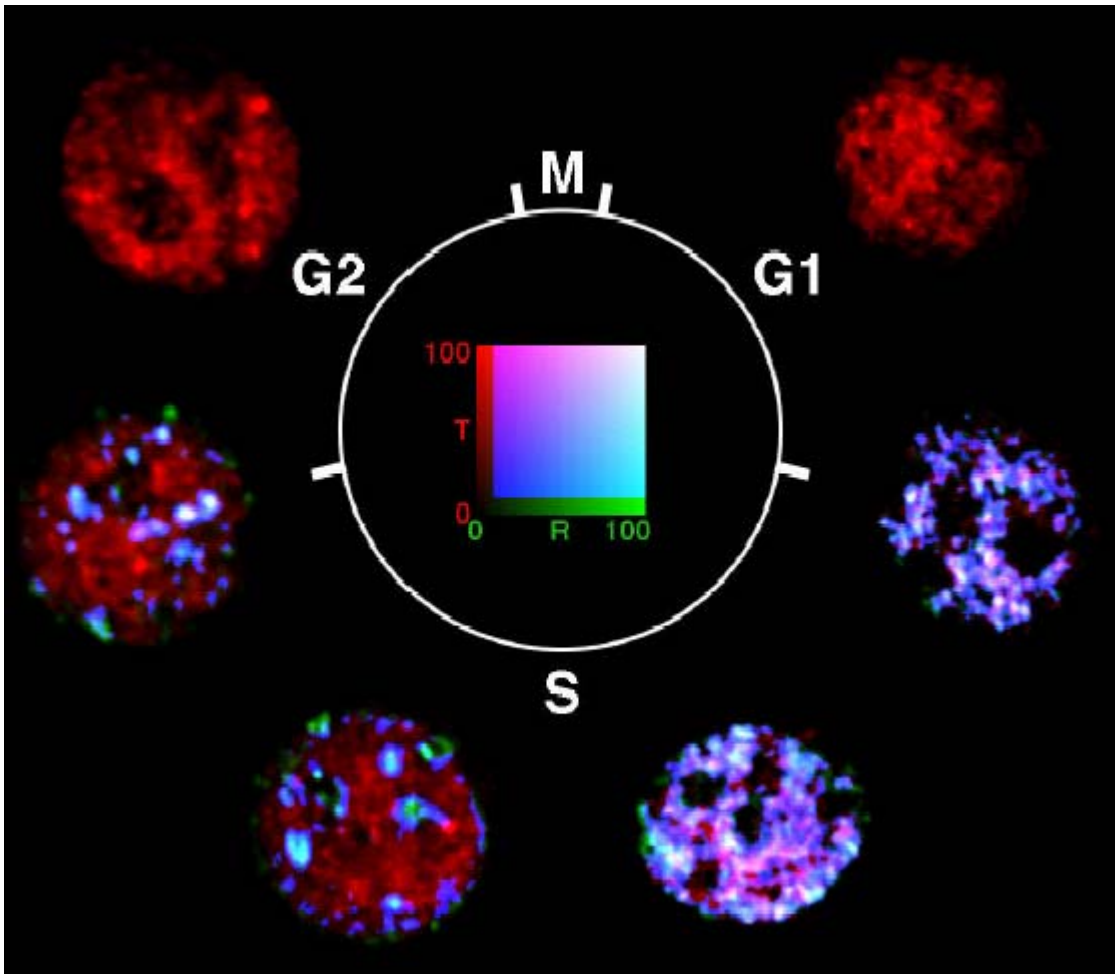

Figure 1.17: Co-localisation of DNA that is replicated in early S phase and transcriptional activity. For details see text (taken from (Cook 1999)).

The above studies have been restricted to the few loci that have been assayed. Further studies have addressed the question whether a correlation between gene transcription and replication timing observed at individual loci is observed at a global level across the entire genome using microarray technology, described in section 1.6.2. The study of the *Saccharomyces cerevisiae* genome in this way revealed no relationship between transcription and replication timing (Raghuraman, Winzeler et al. 2001). However, when a similar study was performed on the genome of the higher eukaryote *Drosophila melanogaster,* a correlation between replication timing and the probability of gene expression was found (Schubeler, Scalzo et al. 2002). If a gene was located within an early replicating region of the genome it is more likely to be expressed. This

relationship was found to be highly significant ($p = 10^{-48}$), but the correlation observed was not absolute; 20% of the earliest replicating genes are found to be transcriptionally silent, and conversely more than 20% of late replicating genes were found to be expressed in the cell line used for the study.

Replication timing of the genome is an important level of organisation in the eukaryote nucleus. The chromatin in higher eukaryotes is not as mobile as the chromatin in yeast (Gilbert 2002). As a result it is important that the transcriptionally active chromatin is available to the transcription machinery. The early replicating DNA must also be available to have access to the replication enzymes. As a result the early replicating, transcriptionally active DNA is found associated with open euchromatin. The question remains whether early replicating DNA leads to transcriptionally active chromatin, or vice-versa. Early replication is currently thought to be necessary, but not sufficient for gene transcription.

Reported genes injected into early S phase nuclei were found to be more than ten times more transcriptionally active than the same gene injected into late replicating S phase nuclei (Zhang, Xu et al. 2002). These transcriptional states remain stable as the cell continues cycling. The reporter genes injected into the early replicating DNA were packaged into chromatin containing acetylated histones, whereas late-injected genes were hypoacetylated (Zhang, Xu et al. 2002), (Goren and Cedar 2003). This reveals that the acetylation status of a region is correlated to the time of its replication.

Two models have been proposed to explain the relationship between early replication and transcriptional activity. Model 1 (Figure 1.18a) suggests that transcriptional potential is established by the early replication of DNA. During S phase, specific activating proteins are available for incorporation into the chromatin, aiding transcription, conversely during late S phase the proteins that are available produce transcriptional regression (Gilbert 2002), (Gilbert 1986). The second model (Figure 1.18b) suggests that closed, transcriptionally-inert chromatin and heterochromatin postpone the commencement of replication by confining the access of replication proteins to the chromatin in the early part of S phase (Gilbert 2002), (Stevenson and Gottschling 1999).

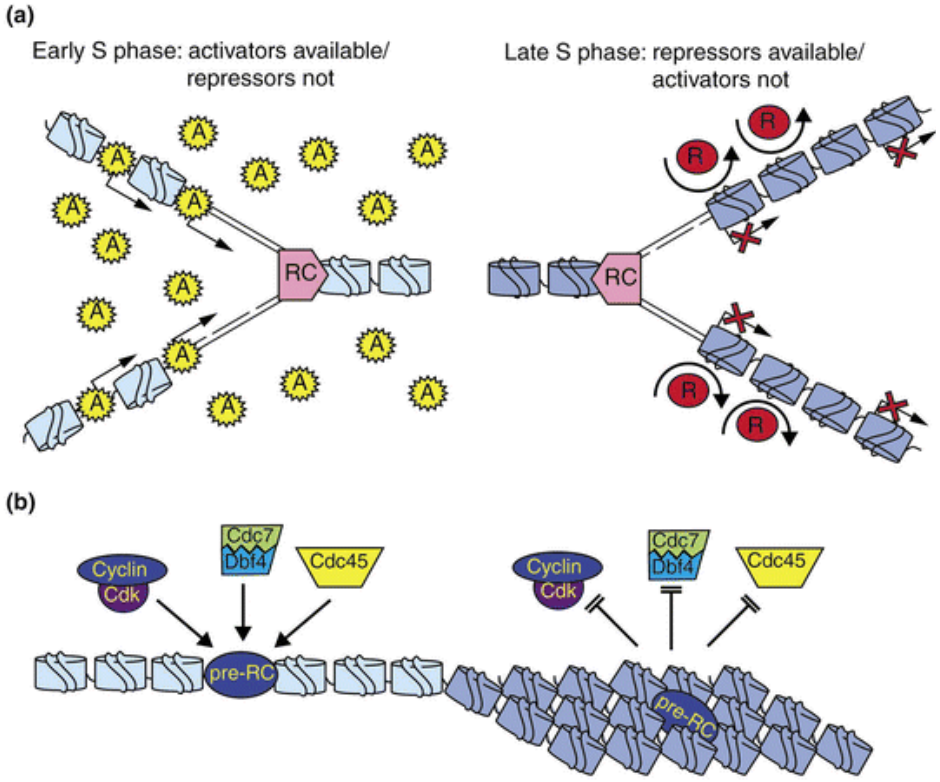

Figure 1.18: Models for linking transcription and replication. A: transcriptional activators are recruited to the chromatin during early S phase, conversely transcriptional repressors are recruited to the chromatin during late S phase. B: Replication initiators can access open, transcriptionally active chromatin first, in early S phase, whilst initiation of origins in heterochromatin is delayed due to inaccessibility of the replication origins (Gilbert 2002).

These models are not mutually exclusive, and data has been found to support both. The first model is supported by the work of Rountree et al (Rountree, Bachman et al. 2000). They illustrate that transcriptional repressors are recruited into the chromatin during late (but not early) S phase. DNA methyltransferase localises to sites of nascent DNA replication in late S phase and recruits histone deacetylases. The deacetylation of the chromatin promotes transcriptional repression.

The second model is supported by late replicating telomeres in yeast. Stevenson *et al* (Stevenson and Gottschling 1999) showed that the condensed telomeric DNA has an inhibitory affect on the replication timing of this region via the Sir-3 protein, which also has an affect on suppressing transcription at the telomeres.

The co-localisation of replication factories and transcriptional machinery in the interphase nuclei may also account for the correlation between early replication and transcriptional activity. Conventional thinking suggests that RNA and DNA polymerases move along the stationary DNA during replication and transcription. However experiments in rat fibroblasts in the S phase of the cell cycle show that the incorporation of BrdU into nascent DNA is not distributed evenly throughout the nucleus, but seems to be focused in approximately 150 distinct sites (Nakamura 1986), (Cook 1999). This supports results described in section 1.3.3, which also show replication foci in early S phase (Ferreira, Paolella et al. 1997; Schermelleh, Solovei et al. 2001). Evidence suggests that the newly synthesised DNA is attached to the nuclear matrix as these foci remain when chromatin is removed (Hassan and Cook 1993). Electron microscopy confirmed the attachment of large DNA replication factories to a diffuse nuclear cytoskeleton (Hozak, Hassan et al. 1993). Similar experiments revealed that RNA polymerases are also attached to the nuclear matrix (Hozak, Jackson et al. 1994), (Cook 1999).

In HeLa cells, fluorescence studies showed a near perfect overlap of sites of transcription and sites of replication (Hassan, Errington et al. 1994) Figure 1.17). One model proposed to describe the coincidence of transcription and replication factories on the nuclear cytoskeleton suggests that during the G1 phase of the cell cycle a replication factory is assembled around two transcription factories (Figure 1.19). This may bring an origin close to a factory so that it can attach and permit replication. In this way the DNA that is close to actively transcribed genes is already located close to replication factories so this DNA is synthesised early ((Hassan and Cook 1994), (Cook 1994; Cook 1995)).

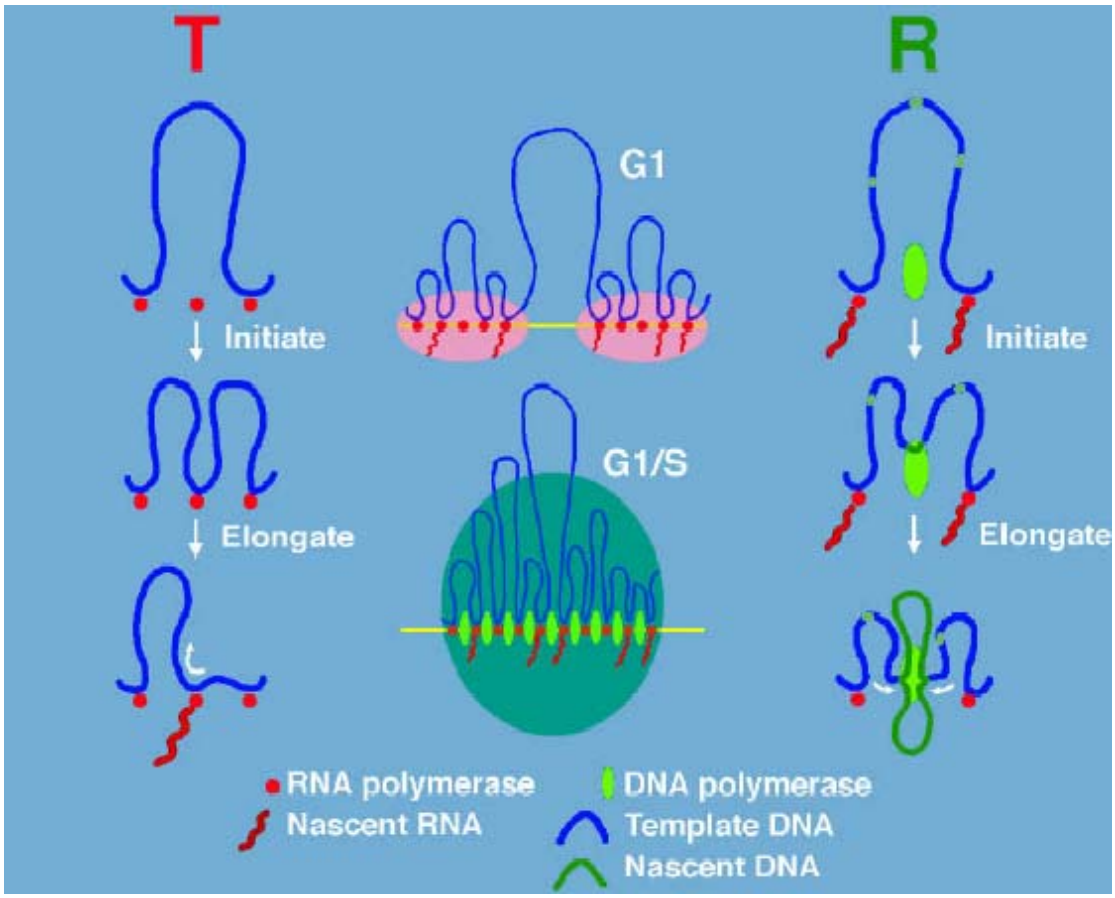

Figure 1.19: The initiation of replication from transcription factories. During transcription (left) DNA is looped to attach to DNA polymerases. The template slides past the fixed enzyme factory to continue transcription. Two of these transcription factories come together to form a replication focus at the G1/S phase boundary. DNA replication is then initiated as an origin of replication binds to a DNA polymerase (right). Template DNA slides along the DNA polymerase as loops of semi-conservative newly replicated DNA are formed (Cook 1994; Cook 1995).

From these studies, many questions remained unanswered concerning the association between replication timing and gene expression. For instance it is unknown if early replication drives the transcriptional activity of a region or vice versa. It is clear however, that other epigenetic factors have a role in the coincidence of early replication and transcriptional activity. The role of these epigenetic factors at sites of replication and transcription will have to be determined before it is possible to understand how DNA replication and RNA transcription interact.

1.6: Using Genomic Arrays to investigate Copy Number Changes.

As stated in section 1.3, an ideal way to assay replication timing would be to assay large regions with high accuracy. In this thesis I describe how replication timing can be assayed using genomic microarrays using their ability to quantify copy number differences between a test and a reference sample, as described below.

1.6.1: Using Genomic Arrays to investigate Chromosomal Copy Number Changes.

Comparative genomic hybridisation was developed in 1992 as a way to detect DNA copy number changes in DNA samples (Kallioniemi, Kallioniemi et al. 1992). The principle of this procedure is co-hybridisation of differentially labelled test and reference DNA onto normal metaphase target chromosomes and measurement of the test to reference fluorescence ratio along the chromosomes. Deviation from the 1:1 ratio of the intensities of test: reference DNA indicated either a copy number gain or copy number loss in the test DNA. However, the detection of copy number changes was limited to the resolution of the signals on the metaphase chromosomes. Low copy number gains and losses can only be resolved if larger than 3-5Mb (Kallioniemi, Kallioniemi et al. 1992).

This principle has more recently been combined with microarray technology to detect copy number changes at a higher resolution (Solinas-Toldo, Lampel et al. 1997; Pinkel, Segraves et al. 1998; Albertson, Ylstra et al. 2000). The metaphase chromosome targets are substituted with nucleic acid target sequences spotted in an array onto a glass slide. The target sequences are obtained from mapped and cloned DNA. As a result, the resolution of the arrays is only limited by the size of the clone and the number of clones represented on the array. The size of the targets that were originally used ranged from 40Kb for cosmid clones, to a maximum of 300Kb for bacterial artificial chromosomes (BACs) (Solinas-Toldo, Lampel et al. 1997). A genome-wide scanning array has been produced which samples the genome with an average resolution of 1.3Mb (Snijders, Nowak et al. 2001). This has been refined to an average resolution on 1Mb using DNA from BACS, PACS and cosmids (Fiegler, Carr et al. 2003). The advent of microarrays utilising overlapping golden path sequencing clones (IHGSC 2001) have allowed the study of whole chromosome arms

with the resolution limited only to the size of the clones and the extent of their overlap (Buckley, Mantripragada et al. 2002).

Originally, to extract enough DNA from clones to spot onto an array large amounts of clone culture needed to be grown (Pinkel, Segraves et al. 1998), (Solinas-Toldo, Lampel et al. 1997, Albertson, Ylstra et al. 2000). To avoid the logistical problems involved in growing large amounts of culture for each locus on the array, methods have been developed to amplify DNA from small amounts of clone DNA, yet still ensuring that the full sequence within the clone is covered. This has been achieved using rolling circle amplification, linker adapter PCR and by DOP-PCR. Rolling circle PCR amplification of the clone DNA utilises the proof reading phi 29 polymerase (Buckley, Mantripragada et al. 2002). Ligation-mediated PCR which produces representative amplification of the genome from just a single nucleus (Klein, Schmidt-Kittler et al. 1999; Snijders, Nowak et al. 2001). DOP-PCR uses amplification of clone DNA by three different, specifically designed degenerate oligonucleotide primers (DOP). This not only ensures the complete amplification of the clone DNA but also ensure that there is minimum contamination from the *E. coli* host vector DNA (Fiegler, Carr et al. 2003).

One problem with using clones from the golden path is that they can contain a high amount of repetitive DNA sequence. This can lead to cross hybridisation with other regions of the genome. This problem is negated by the inclusion of Cot 1 DNA in the hybridisation mix.

A second problem is the presence of low copy segmental duplications in the DNA represented on the array. This again results in cross hybridisation with other regions in the genome; however, these effects cannot be competed out with Cot 1. One way of resolving this problem is not to use DNA isolated from clones. A strategy using PCR products which eliminate segmentally duplicated and common repeat elements has been implemented which avoided problems caused by cross hybridisation to secondary areas of the genome (Buckley, Mantripragada et al. 2002).

Genomic arrays have been used for many applications, such as the detection of copy number changes in cancers (Albertson, Ylstra et al. 2000; Albertson 2003), congenital

microdeletion studies (Buckley, Mantripragada et al. 2002), cytogenetic chromosome rearrangement (e.g. at chromosomal breakpoints) (Fiegler, Gribble et al. 2003) and epigenetic studies (van Steensel and Henikoff 2003).

1.6.2: Using Genomic Arrays to assess Replication Timing.

The ability of microarrays to detect copy number changes has also been used as a novel method to assess replication timing. The principle is to assess changes in the amount of DNA present at a particular locus during S phase. As the change in DNA copy number is very small (a maximum two fold difference) it is essential that the technique is very precise.

The first organism to have its genome analysed in this way was the yeast *Saccharomyces cerevisiae* (Raghuraman, Winzeler et al. 2001). Figure 1.20 from their publication shows how this was achieved.

Newly replicated DNA was labelled with carbon and nitrogen isotopes; early replicating DNA was labelled with the light C^{12} and N^{14} isotopes (HL) whilst late replicating DNA was labelled with the heavy C^{13} and N^{15} isotopes (HH). After synchronisation of cells with α factor, samples were taken at specific time intervals and a restriction digest was performed. The DNA was fractionated by caesium chloride density centrifugation which separates DNA according to the density of the labels. The two DNA populations were then separately labelled with biotin and individually applied to a high density array containing probes covering the entire *Saccharomyces cerevisiae* genome. At each locus the hybridisation ratios of the separate experiments were compared and plotted against position on each chromosome.

The high density of the *Saccharomyces cerevisiae* array allowed replication origins to be mapped. These appeared as peaks on the replication timing profile. Slight differences in replication timing due to the progression of the replication fork were detected and the replication origins mapped. Fork migration rates were also calculated by determining the slope of the profile around the ori (Figure 1.20c). While the link between replication and transcription in *Saccharomyces cerevisiae* was investigated,

in general no correlation was found. This was puzzling as it had previously been clearly observed in higher eukaryotes. The only exception was a family of histone genes which replicated 10 minutes earlier than the genome average of 31mins after release into S phase.

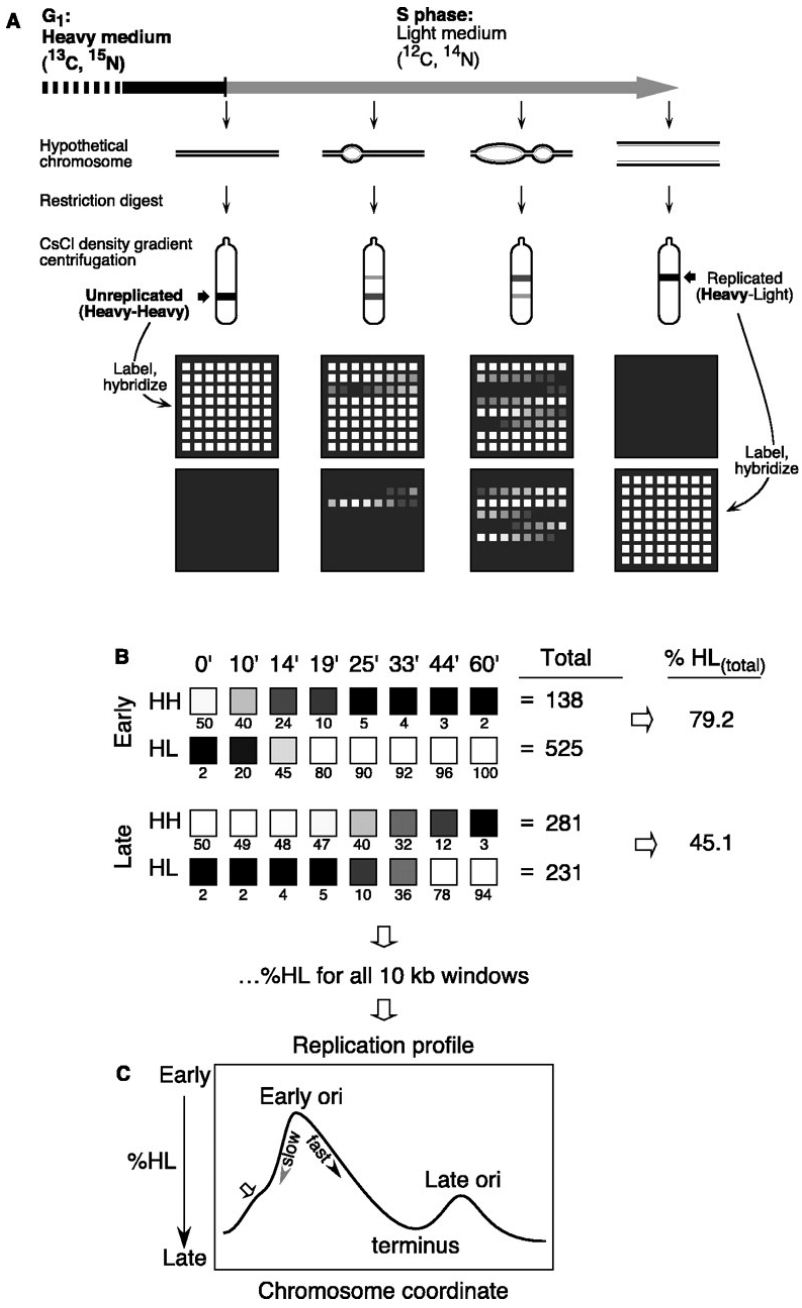

Figure 1.20: From (Raghuraman, Winzeler et al. 2001) illustrating how replication timing was assessed in *Saccharomyces cerevisiae.*

This study paved the way for the investigation of the replication timing of other organisms using microarrays. The measurement of replication timing on an array containing cDNA sequences from *Drosophila melanogaster* (Schubeler, Scalzo et al. 2002) allowed the correlation between replication timing and transcription to be assessed in a higher eukaryote. The method used was slightly different; BrdU was incorporated into the DNA of an unsynchronised *Drosophila* cell line. Flow sorting was then used to separate nuclei in early and late S phase using propidium iodide staining. The newly replicated DNA from each phase was isolated by BrdU immunoprecipitation. The two samples were differentially labelled using Cy3 and Cy5, before being co-hybridised to the same array (Figure 1.21).

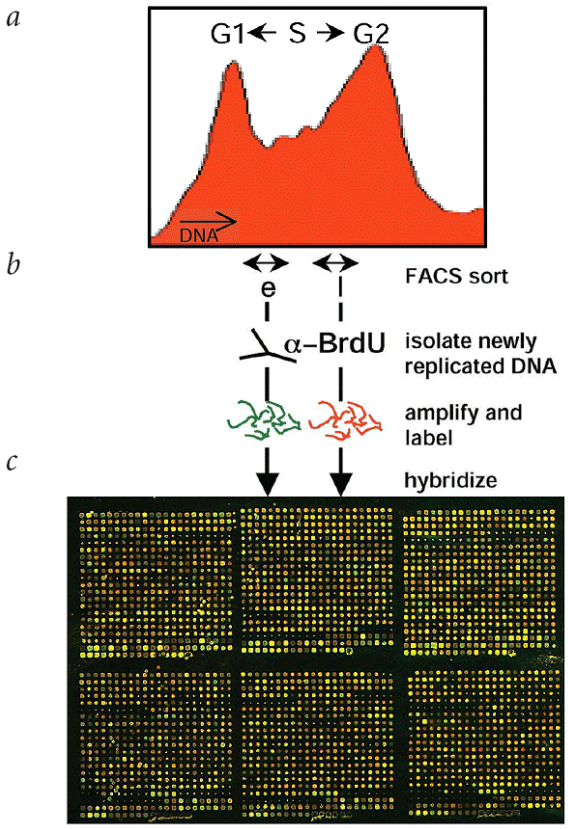

Figure 1.21: Reproduced from Schubeler et al 2002, illustrating how the replication timing of *Drosophila melanogaster* was assessed using microarray technology.

The array used contained 5,221 cDNAs giving an average sampling resolution of 20.5Kb. It also included probes containing retrotransposable elements, which map to blocks of repetitive DNA. The $\log_2$ ratio of late: early DNA was plotted against chromosomal position to produce a replication timing profile for each chromosome,

as shown in Figure 1.22. However, a higher resolution array would need to be used if origins were to be mapped, as in the yeast.

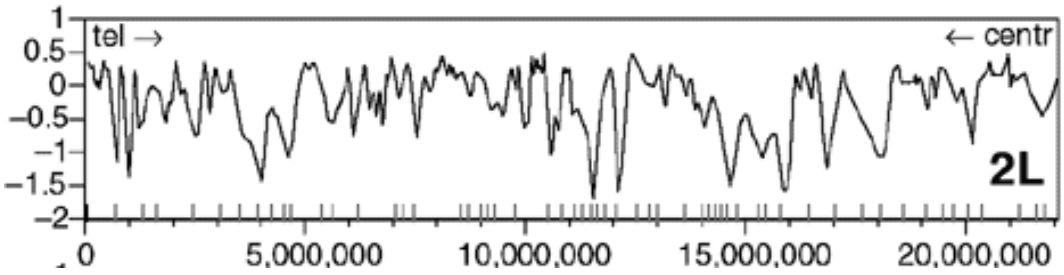

Figure 1.22: Replication profile of *Drosophila* chromosome arm 2L. The $\log_2$ transformation of the data meant that early replicating loci have more positive values, while late replicating regions have more negative values (Schubeler, Scalzo et al. 2002).

The cDNA array described was also used to measure transcriptional activity of the same *Drosophila* cell line. The use of microarray data ensured that there were enough data points to statistically correlate replication and transcription (Gilbert 2002). This study showed a correlation between the replication timing and the probability of gene expression. Early replication coincided with a higher likelihood of the gene being expressed and this correlation was highly significant.

The microarray probes used were derived from cDNA libraries and expressed sequence tags (EST's) which represented less than half the predicted number of *Drosophila* genes. This ensures that the analyses were conducted on coding DNA (McCune and Donaldson 2003). To determine the replication timing of non-coding regulatory regions and to understand how replication timing of non-coding DNA affects other characteristics of the genome such as transcriptional activity and the epigenetic code, any array used must contain representative genomic sequence. The arrays described in this thesis use cloned genomic DNA allowing correlations between replication timing and other features of the genome to be calculated for coding and non-coding DNA. In addition, large scale analysis of replication timing is carried out on a genome wide basis.

1.7 Aims of this Thesis

The main goal of this project was to use genomic microarrays to assess replication timing at a genome-wide level. The resolution of the replication timing map was then refined by the production of arrays from tile path clones.
A chromosome 22q genomic microarray was assembled. This was used to produce a replication timing map for chromosome 22q and to assess other features of the chromosome such as histone acetylation and copy number changes.

The project can be summarised with these aims.

1: A pilot study on a 4.5Mb region of chromosome 22. This involved the production of an array from clone DNA to cover the region chosen, the assessment of replication timing by co-hybridisation of differentially labelled S and G1 phase DNA to the array and the correlation of replication timing with GC content and gene density. (Chapter 3)

2: The production and verification of an array covering the whole of 22q using tilepath clones (average resolution 78Kb); the production of an array covering 4.5Mb of chromosome 22 with 500bp PCR products (average resolution 10Kb); and an array covering 200Kb of chromosome 22 with overlapping 500bp PCR products. (Chapter 4)

3: The production of a replication timing map of the whole genome in a lymphoblastoid cell line. This was performed at a 1Mb resolution and at a tile path resolution for chromosomes where tile path arrays are available. This data was then used for the assessment of the correlation between replication timing and sequence features of the genome; specifically GC content, gene density and density of common repeat elements. (Chapter 5)

4: The assay of the transcriptional activity in the lymphoblastoid cell line and the correlation of transcription with replication timing. The 22q tile path array was used to assess histone acetylation by chromatin immunoprecipitation and application to the

array. The acetylation status of chromosome 22 was correlated with replication timing (Chapter 6).

5: The assessment of microdeletions on chromosome 22 using the 22q tile path array (Chapter 7).