Developing Models of the Mammalian Cell S Phase

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Engineering and Physical Sciences

2010

Alexander George Shaw

School of Chemical Engineering and Analytical Science
# List of Contents

Contents Page | Page
--- | ---
Contents Page | 2
Abbreviations | 4
Abstract | 5
Declaration & Copyright Statement | 6
Acknowledgements | 7
1.0 Introduction | 8
  1.1 Project Aims | 9
  1.2 Project Plan | 11
2.0 Literature Review | 12
  2.1 Introduction | 12
  2.2 Mini Chromosome Maintenance Complexes | 13
    2.2.1 MCM Binding | 14
    2.2.2 MCMs & Single Round Licensing | 17
    2.2.3 MCMs Distribution | 18
  2.3 DNA Replication - Origins and Replication Forks | 22
    2.3.1 Origin Firing | 22
    2.3.2 Origin regulation by Cdk2-Cyclin E | 24
    2.3.3 Distributions of Origin Firing | 25
    2.3.4 Disruption of Origin Firing | 25
    2.3.5 Replication Fork Progression | 26
    2.3.6 Fork Merging | 27
    2.3.7 Fork Stalling, ATM and ATR | 27
    2.3.8 Recovery of Replication Forks | 30
  2.4 DNA Organisation - DNA Foci, Chromosome territories and Chromosomes | 31
    2.4.1 Replicon Clustering | 31
    2.4.2 Structure and Conservation of Replicon Clusters | 34
    2.4.3 Higher level organisation of the Nucleus | 38
    2.4.4 DNA dynamics during Replication | 41
  2.5 Replication Factories | 42
    2.5.1 Structure and Composition of Replication Factories | 42
    2.5.2 Replication Factory Dynamics | 43
  2.6 Spatial and Temporal distributions of DNA replication over S phase | 44
    2.6.1 The Patterns of S Phase Progression | 44
    2.6.2 Why are these patterns seen? | 47
    2.6.3 How can replication be measured over the genome? | 49
  2.7 S Phase Checkpoints | 50
  2.8 Previous modelling relevant to the project | 53
    2.8.1 Modelling of DNA Replication | 53
    2.8.2 Modelling of chromosome structure, territories and DNA. | 57
3.0 Material & Methods | 60
  3.1 BrdU Labelling | 60
  3.2 Biotin Labelling | 61
  3.3 EdU Labelling and staining via Click-IT® | 62
  3.4 DNA Fibre Spreading (Jackson & Pombo, 1998) | 63
  3.5 DNA Fibre Immunostaining | 64
  3.6 Whole Nuclei Immunostaining | 65
  3.7 Serum Starvation Synchronisation | 67
  3.8 Nocodazole Synchronisation | 67
Abbreviations:

MCM - Mini Chromosome Maintenance complexes
Pre-RC - pre-Replicative Complex
ORC - Origin Recognition Complex
ARS - Autonomous Replicating Sequence
PEN - Penta-Nucleotide Palindrome
SV40 - Simian Virus 40
ssDNA - single-stranded DNA
CDK - Cyclin-dependent kinases
RPA - Replication Protein A
PCNA - Proliferating Cell Nuclear Antigen
ATM - Ataxia telangiectasia mutated
ATR - ATM and Rad3-related
DSBs - Double Strand DNA Breaks
R-(Reverse) bands
G-(Giemsa) bands
MRN - Mre11-Rad50-Nbs1
SLs - Small-scale chromatin Loops
GLs - Giant-scale chromatin Loops
FISH - Fluorescence In Situ Hybridization
GFP - Green Fluorescent Protein
I(t) - Defined as the number of initiations per time unit per unit length of DNA
KJMA - Kolmogorov-Johnson-Mehl-Avrami
IEF - Increased Efficiency Model
REM - Relative Efficiency Model
BrdU - 5-Bromo-2-deoxyuridine
Biotin - Biotin-11-dUTP
EdU - 5-ethynyl-2’-deoxyuridine
Abstract

Developing Models of the Mammalian Cell S Phase
Alexander George Shaw, 2010
A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy

The accurate replication of the mammalian genome is a complex and logistically challenging process. The entirety of the genome must undergo a single duplication with as little error as possible. This must occur in a coordinated fashion and over suitably short time scale so as to allow timely cellular division within a cell cycle that is typically around 24 hours in a human cell.

A great wealth of knowledge already exists describing various aspects of the S phase, during which this replication of the genome occurs. This data has been gathered over a variety of model systems, ranging from inferences from the replicative mechanics of SV40 through to direct observations of replication in mammalian cells.

In order to integrate this data and determine the value of inferences from different data sources, quantitative models of the mammalian cell S phase are required. This study documents the development of several such models and the exploration of the influences that experimentally determined parameters and different mechanistic theories can have on the behaviour of a simulated S phase. Of particular exploratory interest were the modes of activating replication of replicon clusters, with the aim of simulating experimentally observed dynamics. Additionally, the study also aimed to investigate the variation of replication fork rates and the density of origins of replication, along with the relationship that occurs between the two during both replicational stress and during a normal S phase.

Through an iterative series of models, relevant parameters and key theories are sequentially explored so as to better understand the S phase. Particularly influential parameters were identified and studied in detail, with experimental determination where necessary in order to more accurately inform the model system. Conclusions concerning the behaviour of the system and the potential impact of the results were drawn upon the completion of each level of modelling and experimental work.

To conclude the study, a linear model simulating the genome of the MRC5 cell line was used to estimate the modes activation of DNA replication along chromosomes in order to recreate experimentally observed replication dynamics. Experimentally determined profiles of replication fork rates and the density of origin firing were also determined for the MRC5 cell line, and were used to populate the model with accurate and appropriate data. Using the model to simulate S phase through a variety of behavioural parameters, realistic S phase dynamics were found to occur through a combination of de novo activation of replicon clusters and a specific probability of neighbour activation by completed clusters.

These derived mechanics, when performed on a system correctly parameterised with suitable data, can simulate experimentally observed phenomena. The development of the model highlighted the requirements of data fit for purpose, and the study also stresses the need for critical consideration of inferences made between different model systems.
Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://www.campus.manchester.ac.uk/medialibrary/policies/intellectual-property.pdf), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on presentation of Theses.
Acknowledgements

I would like to acknowledge and thank Dr Dean A Jackson for his supervision and guidance throughout this study. Guidance was also gratefully received from Dr Apolinar Maya-Mendoza with regard to the majority of the experimental protocols and microscopy. Pedro Olivares-Chauvet kindly processed the TimEx-Seq dataset and aided with formulation of experimental protocols. I would also like to thank the Jackson Group for their feedback and suggestions concerning numerous aspects of the study. I am also grateful for the aid of Dr Kieran Smallbone during data analysis and Dr. Nils Blüthgen for his input in the initial stages of the project.

Finally I would like to thank Robert Platt and Jennifer Withers for their collaboration during the creation of the initial model on which the project was based.

I would also like to acknowledge the EPSRC & BBSRC for funding through the Manchester Doctoral Training Centre for Integrative Systems Biology.
1.0 Introduction

The accurate replication of the mammalian genome is a complex and logistically enormous process. The entirety of the genome must undergo a single duplication with as little error as possible. This must occur in a coordinated fashion and over suitably short time scale so as to allow timely cellular division within a cell cycle that is typically around 24 hours in a human cell.

Unlike lower organisms such as prokaryotes, mammalian cells and other higher organisms replicate their DNA from a series of replication bubbles which begin at potential origin sites across the genome. The firing of origins occurs within replication factories which assemble in association with these sites. Some eukaryotes such as yeast have sequence specific identification of potential origins; mammalian cells, however, show no such correlation. Origin sites are instead likely to be defined by a number of factors including the structure of the chromatin and perhaps transcription.

The occurrence of DNA replication across the genome follows both a spatial and temporal pattern throughout S phase. Replication is seen to begin in areas of euchromatin throughout the more central areas of the nucleus, moving into heterochromatin around the cellular periphery and nucleolus as S phase progresses. Again, as there is no sequence specific basis for this behaviour, the observed dynamics of factories may be dependent on similar factors to potential origin distribution. Whatever the controlling factors, the programmed activation of DNA replication must facilitate the replication of the entire genome and must do so in an efficient manner.

This project will explore the factors that create this programme of S phase through study of its components such as replication factory dynamics, replication fork rates, origin distribution and nuclear architecture. These parameters can then be used to create and augment models of the system in order to evolve new questions and refine the understanding of mammalian DNA replication. A key component of this exploration will be the investigation of the S phase progression, which is likely to have a crucial role in the replicative process.
1.1 Project Aims

In order to approach this topic from a systems biology perspective, a number of specific aims have been explored during a process of iterative modelling. This thesis documents the development of a number of models which become increasingly complex due to the addition of new parameters and concepts, the requirements of which become apparent as each model is evaluated. The keys aims that are developed during this project are outlined below.

1. *Create iterative models of DNA replication on a genome wide scale*

The structure of the project will be provided by an iterative series of models of DNA replication, beginning with a model of replication on a chromosome wide scale. This will act as a benchmark for future models and can be used to extrapolate additional factors required in order to refine modelling and our understanding of DNA replication. As new data is gathered and additional concepts must be considered, new models will be created to incorporate the data and identify further key areas of exploration.

2. *Replication forks and Origin densities*

Several previous studies have been conducted concerning the speed at which replication forks progress during DNA replication (Jackson & Pombo, 1998, Conti et al, 2007). Some of these studies have given conflicting results, suggesting that fork rates may be dependent on a number of conditions that may have been altered as part of the experimental procedure. There is also clear evidence of a relationship between the speed of replication forks and their number, provided through the density of origin firing. A detailed study of the parameters affecting fork rates and origin densities will be performed, giving greater insight to the mechanism that relates these two intra-replicon cluster key factors. Fork rates and origin densities are also key parameters involved in the modelling of replication within replicon clusters, which in turn has a knock-on effect to replication across the genome, making them essential to investigate if we are to create accurate models.
Such a study would involve the immunolabelling of DNA fibres in order to determine the speeds of replication forks and their distribution under a variety of controlled conditions.

3. *Distribution and timing of DNA replication across the nucleus*

DNA replication within the mammalian cell S phase follows a strictly maintained spatio-temporal pattern (see Section 2.6.1). Whilst a number of factors have been implicated as potential guiding influences in constructing this pattern, there is no consensus explanation. The reason(s) that some clusters are preferentially activated over others and the methods by which factories progress from one cluster to the next remain unknown. A combination of nucleus-wide analysis and mathematical modelling will be employed to investigate replication progression, with the aims of determining what guides progression and how factory dynamics relate to the observed patterns.
1.2 Project Plan

As explained in the first project aim, this study aims to develop sequential models of the mammalian cell S phase. After a brief exploratory model to help refine questions that can be answered by this approach, a series of specific aims will be set in order to begin to move towards fulfilling the overall objectives. These will in turn be explored through modelling, experimental simulations or laboratory experiments as appropriate. Each segment of this project is therefore part of an iterative process, with results feeding back into new experimental designs.

As a result of this method of study, each chapter will include a subset of aims, and any additional method information relevant to the experimental approach of that chapter. From the results obtained, conclusions will then be drawn, followed by an evaluation of the impact of the results. This new information and existing knowledge from different experimental sources will be integrated with prior conclusions in order to develop more refined models and expose new areas that could become important focuses for future research.

Prior to the development of the initial exploratory model, a literature review of the mammalian S phase (with contrast to other organisms) has been performed (see Section 2), allowing focusing of questions and experimental techniques. Further chapters will then document the series of models developed, and the experiments, both biological and theoretical, used to gather information to further refine these models.
2.0 Literature Review

2.1 Introduction

The replication of any mammalian genome is a very complex and logistically challenging process. The diploid human genome in particular contains around 6 billion base pairs of DNA which must be accurately duplicated one single time per cell cycle. To do so from a single origin of replication per chromosome would be impossible; not only would this be extremely time-consuming but would also be extremely susceptible to replication fork failures. Replication therefore occurs from as many as 50,000 origins per cell cycle (DePamphilis et al, 2006) which are seeded across the genome. The stretch of DNA that is replicated from a single origin is termed a replicon, and whilst these can vary greatly in length (the majority being between 50 and 300 kbp, Jackson & Pombo (1998)), most studies agree with an average distance between origins of between 100 kb and 150 kb (e.g. 111kb median for keratinocytes (Conti et al, 2007), 144 +/- 66 kbp for HeLa cell line (Jackson & Pombo 1998)). The large variation is indicative of non-uniform origin distribution, with some areas, such as the transition zones between euchromatin and heterochromatin, having low origin densities and thus being more vulnerable to fork stalling and DNA damage. However, given successful progression of replication forks, oncoming forks will merge to give a continuous stretch of replicated DNA. It is estimated that 10-15% of replicons are actively replicating at any one time, leading to an overall time of 8-10 hours for the replication of the complete genome (Jackson & Pombo, 1998).
2.2 Mini Chromosome Maintenance Complexes

The initiation of replication from an origin requires a number of different complexes and occurs in a carefully regulated manner. The overall process of origin activation is often termed “origin firing” and involves polymerase complexes and helicases progressing along the DNA away from the origin. The sites of potential origins are marked by the presence of Mini Chromosome Maintenance complexes (MCMs) which are the focal point of firing. MCMs are both the site of recruitment for necessary factors and are also actively involved in replication by providing helicase activity. These MCMs are thus essential for initiation and elongation during DNA replication in eukaryotic cells (Bailis & Forsburg, 2004). MCMs are loaded by an Origin Recognition Complex (ORC) during G1 phase of the cell cycle (Costa & Blow, 2007), with the six subunits of the MCM complex (Mcm 2-7) forming a hexameric ring around the DNA (Blow & Dutta, 2005). A pair of MCMs is often termed a “pre-replicative complex” (pre-RC) that has the potential to fire under the correct conditions (Costa & Blow, 2007). The two MCM complexes are oppositely orientated and perform helicase activity on one of the two replication forks incorporated in a replication bubble.

The exact number of MCMs that can be recruited around a single origin is still under debate. Hyrien et al (2003) suggest a single ORC complex can recruit as many as 10 MCM pairs; Blow & Dutta (2005) also cite the concept of 5 to 20 MCM pairs being loaded per ORC. However, they do question whether a single ORC can load MCMs as a double hexamer or alternatively the MCM pairs are formed from the combined efforts of two ORCs with different orientations loading an MCM each.

Despite this abundance of MCMs in higher eukaryotes, it is widely accepted that only a small subset of MCMs actually fire, and that firing is not necessarily dependent on the ORC itself. Hyrien et al (2003) review this model, citing experiments in which show that there is preferential activation of MCMs near ORCs as Cdc7 (a factor required for MCM firing) is only
recruited by ORCs (Pasero et al., 1999). However, in higher eukaryotes, recruitment of Cdc7 is MCM dependent and ORC independent, allowing MCMs to activate anywhere (Jares & Blow, 2000). Possible reasons for this overabundance of MCMs will be discussed in Section 2.4.

Whilst the licensing of DNA with MCM complexes sets the maximum number of potential origins, other reasons have been suggested so as to result in the firing of only a subset of origins. The reasons behind this are likely to be both the structure of the DNA itself, with the firing of one MCM (perhaps due to its positioning with regard to replication factories (see Section 2.5)) meaning that the surrounding DNA is in a less favourable position as defined by the torsion length of the DNA. Alternatively, there may be a molecular mechanism, such as the process of “origin interference” (Lebofsky et al., 2006), which would be theorised to lead to localised inhibition within groups of MCMs; as one fires, it inhibits nearby MCMs as their firing becomes unnecessary. If it occurs, origin interference may operate through the ATR pathway (see Section 3.6.1) or through another as yet undetermined process. However, a more simple explanation may simply involve pre-RCs having varying firing probabilities, leading to the stochastic activation of only a subset of complexes. Such a mechanism has been investigated by Blow & Ge (2009) and has been shown to provide close approximations of experimentally derived data.

### 2.2.1 MCM Binding

**Where do MCMs bind?**

The seeding of the MCMs across the genome of higher eukaryotes occurs during G1 phase of the cell cycle and is referred to as “licensing” as it facilitates the replication of DNA local to it. In lower organisms such as *S. cerevisiae*, the exact location of the licensing points is sequence specific such as the Autonomous Replicating Sequence (ARS) common to many budding yeasts. All ARS share a particular 11 bp motif, known as the ARS core consensus sequence (ACS). This short sequence is thought to be the binding location of the ORC (Bell & Stillman, 1992). However, in higher organisms, and even in other yeast species such as fission yeast, the distribution of origins seems dependent on different factors. For example,
Segurado et al (2003) performed a comparative study of sequences at the origins used by the yeast S. pombe and found no consensus. Only a few examples of sequence specific origin definition occur in mammals. Documented cases are the Chinese Hamster DHFR loci, Rhodopsin loci and human rDNA (Hyrien et al, 2003); however, no such sequence occurs at the majority of areas that contain replication origins.

Within mammals, much of the molecular detail of DNA replication itself has been inferred through the study of the operation of the Simian Virus 40 (SV40) in, a double stranded DNA virus, when replicating in Xenopus egg extract. The virus operates through a mechanism which avoids the requirements of DNA licensing, with both the initiation of replication and the helicase activity usually associated with MCMs being performed by the viral Tumour (T) antigen (Stahl et al, 1985; Stillman & Gluzman, 1985). Replication occurs at a specific virally encoded origin which consists of three core regions, all of which are highly conserved in both sequence and spacing (Bullock & Simmons, 1997). The first of these regions is the Early Palindrome, also known as the DNA Unwinding Element, which is melted after T-antigen binding. The second element is a cluster of four GAGGC pentanucleotides collectively known as the Penta-Nucleotide Palindrome (PEN). The PEN acts as a binding site for the T antigen. Finally, the third element consists of a 17 bp A-T rich region that allows DNA bending and may also be potentially involved in DNA melting due to its high A-T content (Bullock & Simmons, 1997). Collectively, these sites mediate the binding of required factors and the melting and unwinding of DNA in preparation for replication. With the addition of a topoisomerase, a single-stranded DNA (ssDNA) binding protein and ATP, the system is independently capable of T-antigen dependent bidirectional unwinding of DNA. The simplicity of the system therefore makes it useful for the study of DNA replication in organisms which would otherwise be too complex, with constituent processes containing too many unknowns.

Whilst SV40 is therefore useful for the study of replication and the requirements of origin formation, it provides little information concerning the nature of licensing and the determination of where MCM loading occurs in higher organisms. Metazoans also lack of the sequence based origin determination of S. Cerevisiae, with the causal factor(s) for licensing
DNA at certain loci being currently unknown in the majority of cases. Many potential factors have been suggested however, such as A+T content, chromatin structure and gene expression. Whilst Segurado et al (2003) found no consensus sequences shared by origins, they did detect A+T rich islands in areas where origins form. A criterion of 75 to 72% A+T depending on length of sequences (500 bp to 1 kb respectively) led to a success rate of 90% in predicting the location of potential origins. There are however a small number of localized origins that have been found in mammalian cells, which often share common features. One such site is the DNA-methyltransferase 1 (dnmt1) loci, which contains a 5' origin of replication. The site contains a 536 bp stretch that has 77% A-T content, at least 19 ATTA and ATTTA nuclear matrix attachment motifs and also a perfect match for the yeast ARS. These features have been shown to be often found at specific DNA replication sites within higher organisms (DePamphilis, 1996), yet the overall requirements are still more complex than those for defining a replication origin in lower organisms. DePamphilis et al (1996) have postulated that this may be due to the variations in chromatin structure and nuclear organization that accompany the increasing complexity of organisms.

**How do MCMs bind?**

The binding of MCMs to ORCs to form the pre-RC is known to require both Cdt1 and Cdc6 (Hyrien et al, 2003). ORC binds to chromatin during the mitosis and early G1, with subsequent loading of Cdc6, shortly followed by Cdt1. This combined complex can then chaperone the binding of the Mcm proteins (DePamphilis et al, 2006). The ORC and Cdc6 also have ATPase activity, and together with the chaperoning function of Cdt1 they may potentially act as a clamp loader, breaking open the MCM complex and allowing it to bind. Alternatively they may help assemble the Mcm proteins directly onto the DNA (Blow & Dutta, 2005).
2.2.2 MCMs & Single Round Licensing

Given the requirements of MCM binding, licensing can be carefully controlled through the regulation of Cdt1 and Cdc6 so that it is restricted to a narrow time frame that only occurs once in the cell cycle.

Cdt1 in particular is very heavily regulated, being ubiquitinated by three different ubiquitin ligases in order to keep its levels low throughout S and G2 phases, not only preventing relicensing but also promoting S phase progression through several pathways that rely on Cdt1 degradation (DePamphilis et al., 2006). In light of this drive towards degradation, a mechanism exists that results in a focused burst of Cdt1 activity during G1. This mechanism operates through the recruitment of geminin by chromatin bound Cdt1 towards G2 and early M phase. Geminin inactivates Cdt1 but also protects it from ubiquitination, resulting in a gradual increase of chromatin bound Cdt1 (DePamphilis et al., 2006). Upon the ubiquitination of geminin as cells enter G1 phase there is a surge of active chromatin bound Cdt1 that can drive MCM licensing. It should be noted that disruption to Cdt1 and/or geminin can easily cause DNA replication abnormalities. Studies by Lin & Dutta (2007) showed that if geminin is inhibited through the use of siRNAs, DNA re-replication can occur within the same S phase. This does however activate a G2/M phase checkpoint due to abnormal DNA structure.

The exact regulation of Cdc6 is still a debateable issue, but a general outline would involve the synthesis of Cdc6 during late G1 and early S phase followed by its elimination during the remainder of the cell cycle. At these times, excess Cdc6 is relocated to the cytosol where it is ubiquitinated and degraded. Higher levels of Cdc6 are maintained during late G1 and early S phase through CDK dependent phosphorylation (DePamphilis et al., 2006), allowing it to perform its role in licensing.

Given this high level of control concerning the initial binding of MCMs, those that do associate with chromatin must also all have been removed by the end of DNA replication so as to
prevent re-replication. MCMs that are not involved in active replication will inevitably encounter replication forks that progress along DNA. The fate of these MCMs is not certain; they may dissociate from the DNA as they meet the fork or be carried along ahead of the fork. The essential result of this, whatever the mechanism, is that they are eliminated from stretches of replicated DNA.

MCMs that form part of active replication complexes must also eventually be removed from DNA. This occurs when replication forks merge, whereupon Mcm 2-7 dissociate from chromatin and are prevented from re-binding to chromatin through a combination of CDKs and geminin (Hyrien et al, 2003).

2.2.3 MCM Distribution

Although MCMs are an essential feature in DNA replication, many studies have shown that the majority of MCMs are unused in any given S phase. MCMs have a more complicated distribution than simply being minimally spread across DNA (which could potentially result in gaps either through fork stalling or poor coverage). However, the exact organisation has been proposed to result from processes of varying degrees of complexity. Lebofsky et al (2006) indicated that MCMs are clustered into licensing groups, where clusters of MCMs are focused around a conserved origin zone. The study supporting these findings focused on an area of 1.5 Mb of DNA assembled from multiple DNA fragments in human primary keratinocytes. They found a series of 13.5 kb+/−5.2 kb “initiation zones” where origins firing in separated S phases were concentrated. Although these areas were on average 40.6 +/- 20.7 kb apart, only 1/3rd were used in any given cell cycle, giving an average distance between origins of 3 times this, resulting in a good approximation of reported replicons sizes. However, the conclusion of multiple MCMs being seeded within a single initiation zone is not necessarily proved via the experimental method provided. The identification of MCM clusters may be due to the combination of origins from several rounds of DNA replication that are in slightly different positions, with a single corresponding MCM in each case (see Fig. 2.1). However, the overall
Spacing of potential origins can be further supported through a study by Maya-Mendoza et al. (2007) in which the addition of caffeine was seen to increase the density of origins by 2-3 fold. Caffeine would therefore appear to inhibit the interference that occurs at the ~40 kb level. Alternatively, if the licensing groups do exist, firing within groups could be prevented by a different mechanism, such as contortional stress. Jun et al (2004) calculated that origins would have to be at least 11 kbp apart in order that both could bend back into a central replication point (see Section 2.4.1).

Additionally, Lebofsky et al (2006) found MCMs to mostly be distributed in intergenic regions, with 66% being completely intergenic and 22% having greater than 90% intergenic content. Many areas were also A and T rich, in agreement with Segurado et al’s (2003) study of origin sequences within the yeast S. pombe.

Given that only one third of licensing zones are used per cell cycle and only one MCM pair within a zone needs to fire, there is obviously a great amount of redundancy within MCM distribution. This has been frequently observed within experimental studies. DePamphilis et al (2006) estimate that 50,000 origins are used per cell cycle with MCMs being 40-100 fold more numerous than this figure (Bailis & Forsburg, 2004).

Fig. 2.1 – Two potential distributions of MCMs leading to the same composite map due to the overlay of multiple trials.
Hyrien et al (2003) set two possible models for MCM distribution. The first would be that MCMs assemble prior to S phase at regular intervals to avoid redundancy altogether. This model would appear to contradict the experimental evidence available and also relies on origin firing being extremely efficient— a single misfire could result in stretches of unreplicated DNA and be potentially lethal.

Their second model involves MCMs being assembled by ORCs at the onset of S phase at random intervals, but rather than minimal licensing the DNA is heavily saturated. ORCs could potentially disassemble and move to other locations after MCM assembly at one licensing site. Gaps can then be avoided at the cost of some MCM redundancy. Excess levels of MCM above the level required for a normal S Phase can be seen S. cerevisiae, humans and Xenopus laevis (Lei et al (1996), Burkhart et al (1995), Edwards et al (2002), respectively).

If this redundant model is to be accepted, it is possible there are other reasons behind the relatively excessive process. A number of solutions have been proposed. Laskey and Madine (2003) suggest that distant MCMs that have not fired do still perform an active role of pumping DNA towards replicative complexes and unwinding it in the process. Alternatively, it has frequently been suggested that excess MCMs have a role in the relief of replicative stress due to fork failure. Taylor (1997) first deduced the firing of additional origins through the experiments which reduced fork rates in CHO cells, yet showed consistent total levels of DNA replication. These findings were confirmed by Anglana et al (2003), with selection of origins being shown to respond to perturbations in the nucleotide pool, thus demonstrating the activation of this compensatory system. Additionally, this reaction to replicative stress can be demonstrated through inhibition of the compensatory system. Woodward et al (2006) performed a study of minimalist MCM licensing in X. laevis by using geminin inhibition of MCM binding to determine the minimum levels possible whilst retaining normal replication kinetics. Under these conditions there was no visible difference in origin spacing between minimal
licensing and the control. However, after slowing replication through the use of aphidicolin and then rescuing with caffeine, only a partial rescue was evident in the minimally licensed cells. Woodward et al confirmed this to be an MCM based rescue by performing the same experiment with a Cdk inhibitor added prior to the aphidicolin which prevented new initiation events. Under these conditions no additional origins were seen to fire in an attempt to rescue the cells from replicative stress.

A similar experiment was also conducted in C. elegans with replicative stress in the form of hydroxyurea (HU) to reduce the dNTP pools. Again, only the minimally licensed condition was adversely affected. Such mechanics have also been observed in human U2OS cells by Ge et al (2007). Reduction of licensing through the use of siRNA targeted against MCM5 lead to cells that were highly sensitive to replicative stress such as from HU and aphidicolin. The use of redundant MCMs to make the system more robust to replicative stress therefore appears to be a conserved mechanism. It should be noted however that the firing of additional origins may not occur through an active response system. As discussed in Section 2.9, the firing of additional origins may explicable through variation in the firing efficiency of origins (Blow & Ge, 2009). Less efficient origins would likely only fire in situations where replicative stress or fork stalling prevented their passive replication by forks emanating from more efficient origins. However, from observation it would therefore appear that secondary origins are programmed to fire in an attempt to rescue replication under stress conditions.
2.3 DNA Replication - Origins and Replication Forks

A key component of this project is the study of the actively replicating areas of DNA. At the smallest scale, this involves the analysis of replication forks which emerge from origins of replication that form at activate MCM complexes. The following section provides a detailed study of the transformation of an MCM complex into an active origin and the subsequent dynamics of replication forks as they progress along the DNA strands.

2.3.1 Origin Firing

With the pre-RCs now assembled onto the chromatin in the form of MCM pairs, the DNA strand is licensed for a single round of replication. The firing of pre-RCs to give active origins is again heavily regulated. The pre-RC is activated by a complex of Cdc7 and Dbf4 in addition to S Phase Cdns (Hyrien et al, 2003). Dbf4 is a regulatory binding protein which activates Cdc7 (Jares et al, 2000). The Cdc7-Dbf4 complex then facilitates the binding of Cdc45 through its phosphorylation. This event is essential for firing as Cdc45 acts as a cofactor for MCM helicase activity (Pacek & Walter, 2004).

Successful firing also relies on the phosphorylation of targets at the origin site by Cdk2-Cyclin E (Shechter et al, 2004, DePamphilis et al, 2006). Cdk2 is regulated through its phosphorylation state; it is activated though dephosphorylation by Cdc25A, with the addition of Cdc25A causing an increase in DNA unwinding and replication. Similarly, a decrease in Cdc25A causes a reduction in both. Cdk2 is inactivated through phosphorylation by Wee1 (Shechter et al, 2004 (DNA Repair)). However, it should also be noted that in a study by Hua et al (1997) where excess Cdk2-Cyclin E was added to the Xenopus system, DNA replication was increasingly inhibited. Hua et al theorised this was due to a reduction in Mcm3 binding; a 2 fold increase in Cdk2-Cyclin E led to little change in Mcm3 binding, but a 4-fold increase led to a 50% reduction in binding. ORC binding was unaffected however, indicating Cdk2-Cyclin E
affects Mcm3 binding between the two steps. Cdk2-Cyclin E can therefore act as both an activator and an inhibitor for origin firing.

In the event of the correct conditions for firing, a limited region of DNA at the origin site is denatured to create an ‘open site’ (Walter & Newport, 2000). The gap must then be extended in both directions along the DNA strand, creating a bi-directional replication bubble. This process is begun by the pair of MCM helicases which widen the site, with the open structure being maintained by Replication Protein A (RPA) which binds to ssDNA. With this template created, DNA polymerase α can bind at each of the forks and begin synthesising a ~12 base RNA primer followed by ~20 DNA base pairs (Johnson & O’Donnell, 2005). The DNA polymerase α is then displaced from the DNA by Replication Factor C, which then facilitates the binding of a clamping protein known as Proliferating Cell Nuclear Antigen (PCNA) which forms as a ring around the DNA. Finally, PCNA stimulates the binding and activation of DNA Polymerase δ which is suited to the replication of the long stretches of primed DNA in the leading DNA strand (Johnson & O’Donnell, 2005). PCNA acts to stabilise DNA Polymerase δ, preventing dissociation from the DNA. PCNA may also be involved in the binding of another DNA polymerase, ε, (Johnson & O’Donnell, 2005), in addition to DNA ligase (Montecucco et al, 1998), which would together perform replication at the lagging DNA strand. The overall process of MCM binding has recently visualised through electron microscopy by Remus et al (2009), who have identified the dual MCM complexes formed around DNA strands. This loading would occur at origins via the breaking and reforming of the ring structure of each MCM complex. The DNA is then seen to lie in a central channel that is evident from 3-dimensional reconstructions of pairs of MCM hexamers. The MCM complexes would then have the ability to slide passively along the DNA whilst first being distributed and later as part of the replicative complex.

The overall process of replication origin structure and firing has again relied heavily on deductions made during SV40 studies. However, rather than an pre-RCs being formed from
the MCM complexes, the structure is instead composed of two hexameric T-antigen complexes facing in opposite directions on the DNA strand (Bullock & Simmons, 1997). Once primed in this manner, active replication can occur through use of many of the host factors including RPA, Topoisomerase-1, DNA polymerase α, PCNA and DNA polymerase δ. These form the completed replicative complex in addition to the T-antigen rings which remains associated with the DNA.

Despite these insights into the structure of replication forks, the final structure of the replication complex is still under debate however. Although the helicase and replication activity are normally closely associated, it is possible to decouple the two. It is therefore uncertain as to whether helicase complex is tethered to the replication complex or whether the helicase complex (which is ahead of the replication complex) simply functions at a rate limiting speed (Walter & Newport, 2000). The decoupling process can be performed through the use of either aphidicolin (Walter & Newport, 2000) or exposure to UV light (Byun et al., 2005).

2.3.2 Origin regulation by Cdk2-Cyclin E

As shown in Section 2.3.1, Cdk2-Cyclin E may seem to have contrary functions as it may act as both an inhibitor and activator of DNA replication. This could be potentially due to the stringent regulation of DNA replication with the cell cycles, with the two modes of action functioning at different times in the cell cycle. For example, Cdk2-Cyclin E can contribute to the prevention re-licensing of DNA after G1 through its inhibition of the binding of Mcm3. Hua et al (1997) studied the cellular concentrations of Cdk2-Cyclin E and found after formation of the nuclear envelope the local concentration of Cdk2-Cyclin E around chromatin increased 200 fold. This would therefore prevent association of Mcm3 except during the short period of the cell cycle where the nuclear envelope was being broken down and reformed. MCM formation would therefore be limited to a single round per cell cycle. The exact effects of Cdk2-Cyclin E would therefore be spatially and temporally co-ordinated, with low levels within the nucleus being required for activation, but excess levels otherwise inhibiting MCM
formation. This system of close regulation of origin firing is also contributed to by other Cdk-cyclin complexes, creating a carefully controlled system that operates within the constraints of the cell cycle. Further examples of these complexes and their influences are discussed in Section 2.7.

### 2.3.3 Distributions of Origin Firing

Given the careful control of origin firing within S phase, efforts have been made to characterise their patterns of firing on a genome-wide scale. Goldar et al (2009) have presented a comparison of origin firing across a number of eukaryotic organisms, and found a strong similarity between the profiles. I(t), defined as the number of initiations per time unit per unit length of DNA, was calculated across the genomes of *S. pombe* and *S. cerevisiae* and over both the *D. melanogaster* 2L chromosome and the *H. sapiens* chromosome 6. Each profile was seen to generally increase to a peak at the midpoint of S phase, with a factor increase of between 10 and 20 from origin firing in the first hour. There are then varied forms of decrease, with few origins firing within the last 30% of S phase. When one considers the reduction in unreplicated DNA as S phase progresses, these profiles would imply an increase in origin firing towards mid S phase (although not as dramatic as a 10 fold increase) followed by a sharp decrease, with only minimal firing of origins during late S phase. Models of DNA replication that operate over the entirety of S phase could potentially use this data for verification of simulated data.

### 2.3.4 Disruption of Origin Firing

The process of origin firing can be perturbed experimentally in a number of ways. The levels of chromatin bound Cdc45 can be increased through the addition of caffeine, leading to increased origin firing (Shechter et al, 2004). In *Xenopus* egg extract an increased rate of initiation can also be achieved through the addition of recombinant Cdk2-Cyclin E and Cdc7-Dbf4 (Shechter et al, 2004). As mentioned previously, Cdk2 activity can also be modulated
through the addition or removal of CdcS25A, causing an increase or decrease in DNA unwinding and replication respectively (Shechter et al, 2004 (DNA Repair)).

2.3.5 Replication Fork Progression

Once an origin has fired, the replication complex (containing the DNA polymerases and PCNA) begins to actively replicate DNA. It is likely that these complexes are static (Blow & Dutta, 2005), fixed to the nuclear scaffolding, with the DNA being spooled through as it is first unwound by helicase activity and then replicated. A key observation in this process is the rate of replication, also termed the fork rate. A single replication origin develops two replication forks, which Kitamura et al (2006) have shown remain associated with each other during replication of the relevant replicon. There is some contention over the rates themselves. Conti et al, 2007 performed studies in human primary keratinocytes, measuring a mean fork rate of 1.46 kbp/min within a large range (0.14 to 11.8 kb/min). However, it should be noted that the standard deviation of the data was only 0.7 kbp/min, with 95% of the data lying between 0 and 3.1 kbp/min. They also found that over 60% of the time, paired forks move at matching speeds and therefore also change speed in a co-ordinated fashion. Jackson & Pombo (1998) also studied the rate of extension using HeLa cells and found that in the first 45 minutes of S phase, replication forks grew at an average rate of 1.7 +/- 0.3 kbp/min with a maximum of 2.3 kbp/min. This shows a close similarity in overall fork speed but the range is greatly reduced. There are also the question of whether fork rates change over the course of S phase or in response to distance to the nearest neighbouring origins. Conti et al (2007) found a positively correlated (R = 0.54; p <= 0.001; N = 219) relationship where distance between origins was proportional to speed with an increase of around 0.8 kbp/min per 100 bp distance. However, a profile of the variations in fork speed over a seven hour period has been created through observations in HeLa cells by Takebayashi et al (2005). They observe a reduction in fork rates towards mid S phase which they relate to the progression of forks through the transition areas between R and G-bands, rather than in response to changes in origin densities.
2.3.6 Fork Merging

In order to replicate the entirety of the genome, it is essential that replication forks move along the entire length of the DNA that they are associated with. Inevitably, and in order to achieve complete coverage, the forks of adjacent replicons must meet. There is then a merging of the replication forks, resulting in the complete separation of the two DNA strands. Such a joining can result in stress along the chromatin due to contorted supercoiling; however, this can be relieved through the action of topoisomerases. Upon completion of replication, the replicative complexes dissociate from the DNA. This is an important factor in preventing re-replication, as another round of licensing must occur before the DNA can replicate again.

The merging of forks occurs at a relatively slower rate than that expected from the rate of replication forks, implying that fork merging may be a rate limiting step for DNA replication (DePamphilis & Wassarman, 1980). This may imply that replicons grouped within a cluster therefore only complete replication when all replicons are ready to merge.

2.3.7 Fork Stalling, ATM and ATR

Despite the many heavily regulated steps in their creation, replication forks do encounter difficulties during replication which can cause them to stall. Potential hazards include DNA damage (in the form of either single and double stranded breaks or lesions caused by free radicals) and dNTP imbalances. Stalling can have different effects depending on its severity; if only the DNA ligase and PCNA dissociate from the replication complex (e.g. through aphidicolin treatment), the complex remains stable and can still be restarted. However, if the MCM complex dissociates during fork elongation the stall is irreversible as the complex cannot be reloaded until the subsequent phase of licensing (Bailis & Forsburg, 2004).

Mammalian cells have two known pathways for detecting DNA damage in an attempt to avoid stalling that could become irreversible. These two pathways are the Ataxia telangiectasia mutated (ATM) and ATM and Rad3-related (ATR). ATM characteristically responds to double
strand DNA breaks (DSBs) and ATR to ssDNA (Shechter et al., 2004). It should also be noted that these pathways are also implicated in regulating origin density; neutralising antibodies to ATM and ATR cause increased nucleotide incorporation through the initiation of excess origins (Shechter et al., 2004). Given the variety of functions that the pathways fulfil it seems likely that they may be linked to a variety of different inputs, and may have different effects depending on the stimulant type, the severity of the signal and perhaps the timing of the cell cycle. Additional factors may also be involved in specific checkpoints, such as p38 which has been implicated in the G2/M checkpoint in response to UV induced DNA damage (Niida & Nakanishi, 2006). These checkpoints will be further discussed in Section 2.7.0.

The ATR Pathway

The ATR pathway is activated by Replication Protein A (RPA) in response to ssDNA (Shechter et al., 2004). RPA binds to and stabilises ssDNA as a heterotrimeric complex of RPA1, RPA2 and RPA3 (Dodson et al., 2004). The ability of RPA to stabilise ssDNA make it essential in the replication, recombination and repair of eukaryotic DNA (Dodson et al., 2004). RPA binding is facilitated by ATRIP (ATR Interacting Protein) which also recruits the ATR complex (Shechter et al., 2004 (DNA repair)). The association of RPA with the ATR pathway makes it a possible target for system perturbations; ATR can cause a checkpoint in response to ssDNA upon the addition of aphidicolin, which can be avoided in Xenopus oocytes through immunodepletion of RPA (Dodson et al., 2004). However, similar experiments involving the depletion of RPA1 by siRNA and induced DNA damage in HeLa cells still resulted in a G2/M checkpoint. Dodson et al (2004) question whether this is due to a crossover between the ATM and ATR pathways or whether RPA levels simply were not reduced enough.

In the event of high levels of ssDNA and therefore RPA binding, the ATR pathway becomes active. Chk1 is one of the downstream effectors and is activated through a Ser 345 phosphorylation in human cells (Shechter et al., 2004). Studies by Shechter et al (2004) have confirmed increased Chk1 activation in response to RPA accumulation and also a reduction in response to caffeine (an inhibitor of the ATR pathway). Again, increasing levels of ssDNA
through addition of aphidicolin can trigger a checkpoint, which is likely to be ATR mediated (Shechter et al, 2004). Active Chk1 functions through phosphorylating Cdc25a, which in turn activates a checkpoint, inhibiting both DNA synthesis and the cell cycle until the damage is repaired. The exact point of interference of the ATR pathway may be dependent on the timing of the cell cycle, activating different checkpoints as appropriate (Shechter et al, 2004, DNA repair). During S phase, Cdc7/Dbf4 can be inhibited through an ATR mediated checkpoint which results in the dissociation of Cdc7 from Dbf4. Cdk2/Cyclin E can also be inhibited by chk1 through Cdk2 phosphorylation and inactivation by Wee1 kinase, a downstream effector of Chk1. Finally, the ATR pathway can also inhibit Cdc2/Cyclin B at the G2/M border so as to prevent mitosis (Morgan, 1995).

Other than its role in stalling forks due to ssDNA, ATR has also been implicated with a number of other functions. In a normal S phase, a combination of the ATR pathway and Chk1 are seen to monitor origin density, coupling defects in synthesis at active forks with the localised firing of latent origins in order to allow recovery of replication (Maya-Mendoza et al, 2007). Shechter et al (2004 (DNA repair)) also suggest a potential role for ATR in stabilising “Fragile Sites”. These are long replication zones with low MCM density, therefore making them particularly vulnerable to damage if replication forks collapse due to replicative stress. However, if the ATR is inhibited, fragile sites develop damage even without replicative stress (Casper et al, 2002).

**The ATM Pathway**

The ATM pathway is characterised by its response to DSBs. The detection of the break itself is through the Mre11-Rad50-Nbs1 (MRN) complex which is recruited to the break point and functions to enhance ATM accumulation and facilitate ATM activation (Lin & Dutta, 2007). MRN therefore acts as a damage sensor for the DSBs and as an amplifier for the ATM pathway. The activated ATM pathway then phosphorylates chk2, which is active during the subsequent cascade of further phosphorylation events (Buscemi et al, 2004).
Dependent on the timing of the cell cycle, Buscemi et al (2004) identify two main effector checkpoints that may become active if there are sufficient DSBs to cause ATM activation. In the case of the cell being in G1, chk2 can contribute to the p53 activation of p21\textsuperscript{waf1}, a CDK inhibitor, causing a G1 checkpoint. Alternatively, if the cell is in S phase, chk2 can phosphorylate cdc25, leading to a G2-S phase checkpoint.

ATM does not appear as vital a process as ATR however; its inhibition does not lead to an automatic checkpoint and loss of cell viability. Despite this, it should also be noted that ATM does autophosphorylate and can activate even without DSBs (Shechter et al, 2004). It may also therefore have a role in other forms of regulation.

2.3.8 Recovery of Replication Forks

In the event of a high levels of fork stalling due to DNA damage, ideally ATM or ATR pathways halt the cell cycle until the damage is repaired, at which point replication can continue. However, in the event of minor localised damage, halting the cell cycle and waiting for DNA repair each time fork stalling occurred would add unnecessary delays to S phase. Also, in the event of dissociation of components of the replication complex it may be impossible to continue with a stalled fork even without DNA damage. In this event, it is possible for nearby pre-RCs to be activated and become new replication origins. Such an occurrence would usually be prevented due to local inhibition of firing caused by actively replicating forks. However, upon stalling it is thought that this inhibition is removed, allowing nearby pre-RCs to activate and recover replication through the same activation process that fired the initial origin.

It should also be noted that at the other extreme, both ATM and ATR have roles in the activation of p53, hence can induce apoptosis in response to excessive DNA damage. ATR has been shown to contribute to apoptosis in such cases, but its presence is not a necessity and a deficiency does not block p53 sensitivity to DNA damage (Nghiem et al, 2002). Similarly,
deficiencies in the ATM pathway do not block p53 induced apoptosis, which conversely can potentially leading to a higher rate of cell death due to inefficiencies in DNA repair (Khanna et al, 2001). Each pathway can therefore work synergistically with p53 in tumour suppressive roles, but their lack of signalling does not promote uncontrolled tumour growth through excessive DNA damage.

2.4 DNA Organisation- DNA Foci, Chromosome territories and Chromosomes

In order to consider the dynamics of DNA replication above the scale of origins and replication forks, we must consider the next-level structuring of the DNA. The genome is not simply a tangle of linear DNA, but is instead localised into specific stable structures termed DNA foci (structurally synonymous with replicon clusters), which form on the scale of ~1 Mb. These form subunits of chromosome territories which in turn are subunits of chromosomes. The following section analyses the steps required to move from individual replicons through the sequential levels of structural organisation.

2.4.1 Replicon Clustering

Despite DNA being divided into separate replication units, i.e. replicons, there are further layers of structure involved the organisation of DNA. Early studies by Nakayasu & Berezney (1989) found that DNA replication was localised within the nucleus into discrete replication granules. Further studies specified these actively replicating areas were not simply isolated origins but were instead clusters of varying numbers of replicons (Jackson & Pombo, 1998). Replicons within these clusters fire at similar times and replicate with similar fork speeds.
These structures are also conserved between cell cycles, as shown by Jackson & Pombo (1998) where double exposures to immunolabelling (at the same time points during separate cell cycles) led to the labelling of the same clusters. Similar observations were also made by Sadoni et al (2004) and Ma et al (1998).

One prominent model of replicon clustering views the structure as a rosette, with the origins of replicons towards the centre, forming the foci (Sadoni et al, 2004, Muller et al, 2004). Replicon clusters are not just convenient blocks of DNA however - they also have an important function during replication. As aforementioned, replicative machinery does not move along the DNA; the DNA is instead spooled through. The machinery is therefore fixed in place, as part of a replication factory (see Section 2.5) that sits on an activated DNA foci (see Fig. 2.2). The choice of which pre-RCs are fired, becoming origins, may therefore be determined by which pre-RCs sit towards the active components of the factory. Equally, in the event of fork stalling and the subsequent lack of origin interference, nearby pre-RCs will also be in a position to fire due to their proximity to the firing factors that could be concentrated within the factory. As replication continues, the loops of the rosette will gradually be spooled through until all the DNA within the cluster has replicated. The factory can then be disassembled and reassembled at another cluster to begin replication anew (Leonhardt et al, 2000). Ma et al (1998) estimate there are approximately 1,000 factories for the 10,000 replication sites in mouse 3T3 fibroblast cells, which, with an average replication time of 1 hour per cluster (Jackson & Pombo, 1998) gives an overall replication time of about 10 hours. Additionally, the number of replicons replicated per factory may also be determined by the number of loops the rosette contains given that an origin can only be fired at the end of the loop that sits at the foci.
An alternative to the rosette model has been suggested by Courbet et al. (2008) which features chromatin loops attached to a form of nuclear matrix. Under normal conditions, the areas of DNA attached to the matrix would be specific regions of high binding affinity, which would in turn be the regions most likely to form origins (see Fig. 2.3). Once replicated, a small number of replication marks would be left on the DNA, perhaps at the termination regions. However, additional sites would also be licensed for the event of replication stress, during which the additional sites could also fire to compensate for slow fork speeds. This would in turn lead to a larger number of origins, and hence more replication marks. However, rather than a return to the original large loop structure, Courbet et al. predict that the increased number of replication marks would cause a larger number of origins to bind to the nuclear matrix during the subsequent G1 phase, thus leading to an increase in origin firing in the next S phase. If the stress was then removed, the increased replication fork rates of the
preferential origins would lead to the passive replication of many of the secondary origins. There would then be a decrease in replication marks, followed by a recovery of the long chromatin loops in the subsequent G1 phase. However, the exact nature of the replication marks has yet to be defined despite their essential nature in the dynamics of this model.

![Chromatin Loop Structure under normal conditions and after stress conditions](image)

Fig. 2.3- Replication origin selection co-ordinated by Nuclear Matrix Adhesion

2.4.2 Structure and Conservation of Replicon Clusters

There have been a number of suggested structures that may be the building blocks of replicon clusters. The grouping of these features, not necessarily into clusters but sharing a similar level of classification, is generally referred to as identifying “chromosome territories” (Cremer et al, 2006). Cremer et al used two structures to classify DNA; small-scale chromatin loops (SLs) of 50-200 kbp and giant-scale chromatin loops (GLs) of 1 to several Mbp. Chromosome territories could be formed of either of these structures or potentially both. A likely structure of a replicon cluster that would concur with the previously explained model (Section 2.4.1) would consist of a number of SLs (each a pair of half replicons between two
origins) organised into a rosette, with separate clusters perhaps linked by GLs. Müller et al (2004) comment on observing such radial-loops during metaphase, but were uncertain as to whether such structures were conserved throughout the cell cycle. It was however pointed out that a random-walk simulation of chromatin organisation could potentially create such a structure, hence the process itself would not be overly complex to orchestrate (Müller et al, 2004). Random-walks are often used to simulate protein-folding, hence its adaptation to chromatin modelling is quite apt. Such a model would characteristically feature random localised exploration around a volume before jumping to a new area, which effectively results in a structure similar to the rosettes linked by GLs.

An alternative model to this form of organisation by structure type is the organisation of DNA through fractal crumpling, resulting in folding of DNA into units of no discrete size. Such organisation has been suggested to explain the observed nuclear reaction kinetics of Bancaud et al (2009). A larger fractal dimension was also observed in euchromatic regions, which would result in less densely packed chromatin than heterochromatic regions. The concept of fractal organisation is further discussed later in this chapter, with consideration of the implication for DNA organisation on a nuclear-wide scale as is suitable for fractal-based architecture.

DNA structure of this level is likely to be more dynamic and complex than the simple division into clusters of replicons however. Not only are there the likely existence of linker stretches of DNA, but chromatin may also undergo temporary localised changes. For example, Gilbert et al (2004) found that chromosome territories were generally more open in areas of high gene density (but not necessarily expression). Müller et al (2004) noted the extended loops within these areas of high gene expression and noted that these loops were also maintained by transcription. Chromatin may therefore undergo extensive changes due to its transcriptional state, which in turn are likely to correspond to the gene density of the local area.
In addition to this flexibility in structure, it is also unlikely that chromatin is rigidly fixed in place; Chakalova et al (2005) comment that chromatin loci are mobile but restricted within a confined volume dependent on the time scale observed. They show that, over 1-2 seconds, loci oscillate within a volume with an average radius of 0.3μm, which is indicative of local tethering. However, over longer time scales, more distant movements can be observed (up to 3μm in 10 minutes has been observed) implying that although usually constrained, loci do have the potential for more distant movement.

Despite their potentially transitive nature, chromosome territories, as structural formations of DNA, are still highly conserved between cell cycles. Hence, although there is the potential for temporary movement, the structure itself remains intact with relatively minor conformational changes. Replicon clusters in particular have been shown to be very stable entities as indicated through the observation of conserved staining after multiple cell cycles (Jackson & Pombo, 1998). Given their key nature to both the replication of DNA and as structural units, the characteristics of replicon clusters are likely to be highly important for the purpose of modelling S phase. Potential parameters of interest are the three-dimensional size and volume of replicon clusters and the degree to which they can potentially overlap or intermingle.

Studies of DNA foci in human cell lines commonly provide a range of size values similar to those observed by Bornfleth et al (1999) in both neuroblastoma and HeLa cell nuclei. They observed foci radii in the range of 200-300 nm, with a mean of 250 nm. These measurements were taken across each of the stages of S Phase. However, it is notable that the size of foci vary greatly over the course of S phase, as is both evident in Fig. 2.4 of section 2.6.1 and noted in studies such as Leonhardt et al (2000). An explanation for this may be found in a more recent study by Koberna et al (2005). Through higher resolution obtained by electron microscopy, the ‘larger’ foci observed were instead decomposed to give a more consistent foci size throughout S phase. Average foci diameters for early, mid and late S phase were
determined to be 110nm, 120nm, and 110nm respectively. One would expect the larger foci formations seen in late S phase to be the results of the compact and clustered nature of heterochromatic regions which previous could not be decomposed into constituent foci. These observations were confirmed by Baddeley et al (2010) through the use to super-resolution light microscopy, with observation of an average foci diameter of 125 nm throughout S phase, with a standard deviations of 31.1-32.2 nm.

A study of the overlap between DNA foci in interphase cells has also been performed by Zink et al (1999). Their study focused on the potential for overlap between foci within the same chromosome territory and focused on territories of chromosome 13 and 15. Their experiments involved two 2 hour pulse labels of human fibroblast cells with a four hour chase followed by growth of cells through several cell cycles. Immunostaining was then performed to identify DNA foci, and a subset of chromosomal territories selected for analysis according to the presence of both labels and their identification via FISH. Each territory was also selected only if it avoided contact with neighbouring territories. Overlap within territories was observed to range from 0 to 15% in the majority of cases with a median of 5% for chromosome 13 territories and 10% for chromosome 15 territories. However, the experiment is potentially limited by both the optical limitations and the small number of chromosomes represented. Additionally, the overlap of foci may alter within S phase due to both their change in size (due to increased DNA content) and their movement during replication. The authors did note that overlap of chromosome territories within both G0 and G1 did not differ however, despite the increased density of DNA. This may not hold true during the dynamic processes of S phase however.

Finally, we must also consider the potential for overlap between elements of different chromosomes. This observation may be a key factor if replication is seen to progress spatially across the nucleus rather than linearly along chromosomes. Current studies, however, agree to either a very low level of contact or no intermingling between chromosomes due to the presence of interchromatin space. Such observations were made by Visser et al (2000) for interphase chromosomes. However, the formation of loops of chromatin during transcription
has led to the conclusion that the mingling of chromosomes may occur in response to gene activation. In a study by Branco et al (2008), the overlap of a subset of chromosomes in both resting and active human lymphocytes was measured. The levels of intermingling changed in response to activation (and hence altered transcription). The absolute amount of overlap in resting cells also varies greatly between chromosomes however, from 1-20% of their total volume. It should be noted that the estimated error is up to 80% of the overall values. Low levels of intermingling of chromosomes therefore seem likely on the periphery, but the majority of the chromosomal content remains segregated.

2.4.3 Higher level organisation of the Nucleus

To extend modelling of DNA replication to nucleus-wide simulations, one must also consider the structure of chromatin on a chromosomal and nuclear-wide. The positioning and organisation of chromosomes are likely to have strong influences on DNA replication both due to their spatial arrangement and their chromatin content. Chromosomes exist in greatly varying sizes and volumes, with a range of content such as gene density and R and G banding.

A huge volume of information exists concerning the content of chromosomes with regard to a wide range of factors. A number of these factors have been implicated in influencing DNA replication and these often have distributions that are far from uniform across chromosomes. However, not only do we lack precise knowledge concerning which factors are relevant (and potentially dominant), but many are themselves functionally related. In addition, whilst the information concerning these factors often exists in a linear form, the distribution of these linear records within the nucleus as a 3D structure will also require careful consideration. Both chromosomal positioning and their orientation are likely to therefore be of concern.

It has been known for some time that chromosomes occupy preferred locations within the nucleus and that sub-elements (such as chromosomal arms) can then display characteristic
orientations. Observations by Croft et al (1999) indicated conserved positioning of chromosomes 18 and 19 in relation to each other within human primary lymphocytes. Chromosome 19 was seen to be more associated with the interior of the nucleus, whilst chromosome 18 was closer to the periphery. Additionally, it was noted that chromosome 18 occupied a smaller volume than chromosome 19, despite its larger physical size. However, much of chromosome 18 is G-band in nature and gene poor, hence is likely to be condensed.

However, these observations lead to the question as to whether chromosome positioning is guided by the size of the chromosomes, gene content or both. A further study was conducted by Boyle et al (2001), who extended the study beyond chromosome 18 and 19 to include the entire nuclear content. They confirmed that higher gene density led to preferential location towards to the nuclear interior. However, no simple correlation with chromosome size was noted upon the comparison of locations of similar sized chromosomes. A more extensive survey was again suggested however, with measurements required concerning the location of chromosomal sub-domains. A contrasting study has however been performed by Cremer et al (2001), who concluded that chromosome territories of small chromosomes were located closer to the nuclear centre, commenting that the protocol of Boyle et al led to disruption of nuclear architecture. Whilst debate on the subject continues, it would appear that gene rich regions of chromosomes generally tend towards the nuclear centre, and smaller chromosomes, or at least elements of their chromosome territories, also tend towards the nuclear centre.

With regard to the potential for modelling, these results cumulatively support the expectations of the S phase progression patterns. The clumping of gene rich areas towards the centre of the nucleus would imply the early replication of the nuclear centre, followed by the periphery as the cell progresses into mid S phase. However, the combined controversy concerning chromosome location, along with the potential for variation both between cells and during different time points of the cell cycle may mean that spatial modelling of
replication along chromosomes in 3D can only be approximated until reliable data is available.

Finally, the DNA organisation across the nucleus as a whole must be considered. Lieberman-Aiden et al (2009) recently published an analysis of the areas of interaction within the nucleus by identifying loci that shared close spatial proximity and were at least 20 kb apart. Analysis of the data revealed that the interactions between loci for each chromosome could be broken down into two sets with enriched interactions within the set and depleted interactions between the sets. This would imply the presence of two spatial compartments of chromatin. Of these two compartments, the arbitrarily termed Compartment B featured higher interaction frequencies, indicating more dense packing, whilst Compartment A correlated strongly to the presence of genes. Compartment A would therefore likely represent the euchromatic regions of genome, and thus the early replicating regions. Compartment B would theoretically represent the heterochromatic late regions.

Furthermore, the data also indicates a potential model for the structure of the nucleus as a whole. Prior literature has often proposed that the nucleus can be modelled as an ‘equilibrium globule’, featuring densely knotted and tangled DNA. However, an alternative model was suggested by Grosberg et al (1993), which was first termed a ‘crumpled globule’ and later a ‘fractal globule’. Instead of a tangle of DNA, chromatin would instead be arranged as a series of beads on a string and then crumpled together. Comparison of Lieberman-Aiden et al’s data to the two prior models of the nucleus indicated that a ‘fractal globule’ model may be more likely. The modelling process involved in these theories will be further discussed in section 2.8.2.
2.4.4 DNA dynamics during Replication

Whilst the classification of DNA in varying levels of structure appears to give an attractive simplification of the chromatin environment, one must also consider the fluid nature of the nucleus. Although, as shown previously, chromatin is restricted in its movements, there is still varying degrees of motion that may occur on different structural scales. With regard to DNA replication there must be the potential for quite extreme localised restructuring within replicon clusters. DNA must firstly be spooled through factories, and must also be allowed enough movement so as to make internal areas of chromatin accessible to factories. There must also be accommodation of the doubled quantity of DNA as S phase progresses.

DNA is also remodelled on the scale of chromosome territories during S phase. To facilitate continued replication, DNA has been shown to decondense at chromatin borders in order to expose internal regions allowing the potential for replication (Cremer et al, 2006). Larger scale movements have been observed by Probst & Almouzni (2008) whose study of perichromatin (which replicates mid-S phase) demonstrated that DNA was pulled from the centre of territories, replicated and then re-internalised. This study also implies that both copies of the DNA occupy a similar region to that of the original DNA strand. Sadoni et al (2004) support this view, stating that only local rearrangements occur rather than a complete redistribution of DNA.

Whilst DNA has the potential for quite significant movement during S phase, the resulting distribution is unlikely to be that different from the original configuration, despite the additional volume of DNA. This has important implications for modelling of replication patterns, as DNA foci are therefore relatively static (on a nucleus wide scale) throughout S phase.
2.5 Replication Factories

Replication Factories are the nuclear compartments where DNA replication occurs, and as such are thought to contain the necessary factors, concentrating them around active replication centres. They assemble in association with replicon clusters, spooling through the outstretched loops of DNA, replicating it and then extruding the dual copies.

2.5.1 Structure and Composition of Replication Factories

The composition of replication factories is currently unknown, although there has been much study of their dynamics during S phase. Due to this lack of knowledge concerning their structure, PCNA is often used as a marker of factories (e.g. via GFP fusion). Studies of actively replicating clusters have shown an initial localisation of PCNA adjacent to the sites of nascent DNA formation (Sporbert et al., 2002). The DNA then undergoes local chromatin rearrangements in order to accommodate the doubling of the DNA within the region, with 95% of relative positional changes of DNA foci (during and after replication) not exceeding 0.5µm (Sadoni et al., 2004). Upon completion of replication the PCNA is then observed to move away from the replicon cluster over time (Sporbert et al., 2002). As clusters have been shown to be relatively fixed, this implies that the factories, or at least some of their components, do move. PCNA has proved to be a reliable marker for factories as it is conserved in cells from yeast through to mammals as it is essential for DNA replication (Waseem et al., 1992). Its fusion with GFP also does not appear to interfere with cells in any observed manner whilst at low expression levels (Leonhardt et al., 2000). Further studies by Kisielewska et al (2005) have also shown PCNA-GFP does not affect cell viability, although its effects on the S Phase behaviour in particular were not investigated.

Using PCNA as a marker, it is possible to follow the dynamics of factories throughout S phase. Hozák et al (1993) first observed factories in early S phase, measuring them to be ovoids with an average size of 157x185 nm. Their subsequent study (Hozák et al., 1994) extended the observations to mid S phase, whereupon the factories begin to reduce in number but become
larger, creating structures 250 to 400 nm on their longest axis. Larger factories (or associated structures) then form as S phase progresses, first forming multi-lobed structures and then becoming sausage shaped, up to 700 nm in length by late S phase.

The exact number of factories is still debated, though attempts have been made to estimate this parameter by counting active replication sites. However, this has given a wide range of estimates, from 150 sites in a chick nucleus by Cossmann et al (2000) to over 1,000 sites in the early S phase of a mammalian cell by Ma et al, (1998). A mathematical estimate would lie between these figures; assuming an average fork speed of 1.7 kb/min and an average of 4 origins per replication cluster (each with two forks) (Jackson & Pombo, 1998), an average cluster will replicate 0.82 Mb of DNA in an hour. Given a diploid human genome of around 6 billion base pairs and a 10hr time period to replicate within, this would then require the maintenance of around 740 factories on average throughout the 10hrs.

2.5.2 Replication Factory Dynamics

Given that there are a finite number of factories available to replicate a genome it is important that they have the ability to be transferred in order to access unreplicated replicon clusters. The exact form of this process is however poorly understood. Leonhardt et al (2000) have demonstrated that factories do not remain intact as they relocate; they instead disassemble and then reassemble at the next replicon cluster. The choice of which cluster is “next” is also currently unexplained. Factories may randomly assemble at nearby foci, replicating it if it is required and otherwise moving again. If this is the case, one must also question whether the factory moves in 3 dimensions, or may randomly jump further down the same DNA strand. Alternatively the factory may relocate sequentially down the same DNA strand, perhaps following one of the still potentially-active replication forks at the periphery of the cluster (as the fork will have no opposing fork to merge with). The exterior forks could also act to cause a “domino” effect (Sporbert et al, 2002), with the fork spooling in towards another cluster and triggering the assembly of a factory around it. This model does not
account for jumping intervening G bands however (see Section 2.6.2), or for moving to any other areas of non-contiguous DNA.

2.6 Spatial and Temporal distributions of DNA replication over S phase

Given the high degree of regulation of origin firing and fork progression, it is unsurprising that the spatial and temporal dynamics of replicon cluster firing also follows a highly organised pattern. Exactly how and when this pattern is dictated is still under debate. If the patterns are dynamically set each S phase, an early a time as early G1 phase has been suggested for setting the temporal aspect (i.e. when does an area of DNA replicate during S phase), independent of exact origin specification (Chakalova et al, 2005). Alternatively, factors such as the nature of the chromatin itself may control the observed patterns, in which case replication dynamics are innately programmed into DNA itself.

2.6.1 The Patterns of S Phase Progression

Through techniques such as nucleotide and PCNA labelling, it is possible to visualise the dynamics of factories and hence the occurrence of DNA replication as S phase progresses. Cells labelled in such a way can often be divided into three types (Nakayasu & Berezney, 1989; Ma et al, 1998) which correlate to the progression through S phase.
The classification held by Ma et al (1998) is as follows:

<table>
<thead>
<tr>
<th>Type</th>
<th>S Phase timepoint</th>
<th>Replication Spatial Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Early to Mid</td>
<td>Predominantly in extra nucleolar euchromatic regions</td>
</tr>
<tr>
<td>2</td>
<td>Mid to late</td>
<td>Perinucleolar and perinucleolar heterochromatic regions.</td>
</tr>
<tr>
<td>3</td>
<td>Very late</td>
<td>Satellite DNA heterochromatic regions</td>
</tr>
</tbody>
</table>

A diagrammatic form of each of these phases is shown in Fig. 2.4.

Nuclei stained via EdU protocol (see Methods 3.3) to illustrate S phase replication patterns:

Cartoon of expected staining pattern:

![Early S Phase](image1.png)

![Mid S Phase](image2.png)

![Late S Phase](image3.png)

Fig. 2.4 – Replication Patterns displayed during the eukaryotic S phase.

In studies by Leonhardt et al (2000) the durations of each of these phases were estimated in C2C12 mouse myoblast cells through use of GFP-PCNA to label the replication foci. The estimated duration of the early S phase was 5 hours, with the cell progressing from mid to late S phase after another 3 hours and then remaining in late S phase for another 4 hours.
They do make note, however, that the replication patterns described are best identified at the mid plane of the cell, with patterns looking different through other segments of the cell. This could be one of the contributing factors to discrepancies concerning the identification of replication patterns. An alternative classification has been described by O’Keefe et al (1992) using both synchronised CHO and HeLa cells to provide an accurate time plan through BrdU labelling. CHO cells synchronised through serum deprivation and hydroxyurea blocking were used for the majority of the study, and upon release from blocking their replication patterns were analysed every hour. Observation of the first pattern peaked after only 0.5 hours, which was identified as the replication of areas of euchromatin at many small sites. The second pattern peaked after 2 hours, characterized by larger and more discrete sites of replication. The distribution of these sites begins to move away from the interior and is localized more towards the cell periphery, including some heterochromatin areas. After 5 hours from hydroxyurea release, pattern 3 is seen to peak, with replication restricted to the nuclear periphery and the perinucleolar regions. Replication occurs in a mixture of large patches and some smaller discrete areas, whilst replication towards the nuclear interior is restricted almost exclusively to the nucleolar surface. Pattern 3 could therefore be seen to be the equivalent to the Mid S phase classification of Ma et al (1998), occurring at a similar time point to that estimated by Leonhardt et al (2000). O’Keefe et al’s fourth pattern is seen to peak 7 hours after release from hydroxyurea, with replication sites becoming larger in size and fewer in number. These sites occur throughout both the nuclear interior and periphery, and were identified as heterochromatic regions through electron microscopy. Finally, Pattern 5 peaks at 9 hours after release, with replication being restricted to even fewer and larger sites. The sites are again heterochromatic, with sites towards the interior of the nucleus tending to be larger. Patterns 4 and 5 would therefore likely represent the movement of replication from mid S phase to late S phase and the S phase’s subsequent completion. In the continuation of their study, O’Keefe et al found similar patterns of replication in their synchronized HeLa cells, and a number of other surveyed human cell lines including MRC5. However, it should be noted that while the various patterns did exhibit defined peaks, there were often large overlaps between the different patterns. For example, at the 2 hour interval, 18% of the population exhibited pattern 1, another 18% exhibited pattern 2 and 64%
exhibited pattern 3. Variation in phase duration is therefore clearly evident despite prior synchronization of cells.

2.6.2 Why are these patterns seen?

A wide variety of factors may be responsible for these observed patterns, including chromatin structure, R/G-banding, C/G content and perhaps transcription. However, as many of these overlap (for example, gene rich areas often have high C/G content and R-bands are often gene rich), correlating the correct reason to the observed behaviour is difficult.

R/G-banding has been frequently noted as a close approximation for predicting replication dynamics, with R-bands replicating first followed by G-bands (Jackson & Pombo, 1998). A transition between the two normally occurs at around 4 hours into S phase.

The bands originate from the use of Giemsa staining to cytogenetically identify human chromosomes. R-(Reverse) bands have light staining and are characterised by their early replication and high gene and CpG island density. They tend to contain large numbers of housekeeping and tissue specific genes and are enriched with short interspersed repetitive elements. G- (Giemsa) bands are more darkly stained and are characterised by late replication, low numbers of tissue-specific genes and are enriched for long interspersed repetitive elements (Strehl et al, 1997).

The greatest problem with S phase progression being orchestrated by R/G-banding is that the R/G- transition is often seen as being a definite cut in S phase (Drouin et al, 1990). This cut may even be accentuated to the point of a pause in replication, although this may be due to checkpoint activation as a result of experimental procedures. Some experimental evidence does reason against this sharp interruption however. Strehl et al (1997) observed a series of
1-2 Mb of R/G-band boundaries and instead found a gradient of early to late replication. There may therefore be fair correlation between replication timing and R/G-banding, but due to the gradual nature of the banding and correlating factors such as gene content and chromatin structure there may not be a definite mirroring in replication timing.

The kinases Chk1 and Chk2 have been implicated in the maintenance of a temporal division between early and late replication origins. Reductions in either kinase led to the premature firing of late origins (Costa & Blow, 2007; McGowan, 2002). A potential pathway for this mechanism would involve the phosphorylation of Cdc25 by Chk1 or Chk2 in order to prevent Cdc2 activation and hence prevent origin firing. The activity of such complexes as Cyclin A2-Cdk1 have also been implicated in this process (Katsuno et al, 2009), with Chk1 depletion leading to hyperacetylation of Cyclin A2-Cdk1 and abnormal replication in early S phase. The normal activity of the complex involves it being inactive in early S phase, followed by a notable increase in mid and late S phase, leading to the conclusion that it has a role in the regulation of late firing origins.

A recent study by Malyavantham et al (2008) has looked more closely at the spatial correlation between areas of early replication and transcription sites (TS). Their expectation was to find a spatial overlap when the two areas were labelled, given previous studies concerning gene rich R-bands and early replication. Areas of actively replicating DNA were labelled, cells displaying early replication patterns selected and then the sites of replication were compared to the TS which contained labelled RNA transcripts. RNA pol II sites were also separately labelled. Whilst a close spatial correlation was shown between the three, there was little overlap of labelling. Malyavantham et al proposed a model whereby chromatin loops containing highly transcribed genes are extended out into TS. These could then be retracted prior to DNA replication, resulting in a pool of labelled transcripts next to areas of early replicating DNA. Given this theory, the replication patterns occurring would not be directly predictable through location of areas of high active transcription. However, there would be a close spatial correlation between the two. A similar study performed by Sequeira-Mendes et al (2009) also found a close correlation between replication initiation sites and...
transcription sites in mouse embryonic stem cells. 85% the replication sites mapped to transcriptional units, with half of these mapping to promoter regions, with highly efficient origins being associated with CpG island-promoters. A further implication of gene expression in driving DNA replication is demonstrated through the cell-type specific replication profiles in higher order eukaryotes as a result of differentiation, which broadly results from changes in gene expression (Hiratani et al, 2008).

Such observations also illustrate the potential influence of the nuclear matrix in creating the observed patterns of DNA replication. The division of the nuclei into sub-compartments adds an extra potential layer of complexity, with both replication and transcription sites being clustered into higher order domains (Berezney et al, 2005). A further proposal by Berezney et al (2005) is that the co-ordination of DNA replication revolves around the selection of zones for replication and transcription at a given time point, following their later re-zoning as appropriate.

2.6.3 How can replication be measured over the genome?

Given the high level of co-ordination of replication both spatially and temporally, efforts are being made to map the occurrence of replication across the genome with greater accuracy. An accurate map of replication would allow comparison of replication timing to a variety of other genomic features, the concurrence of which could lead to the identification of causal factors that guide the replication programme.

Recent developments in technologies such as microarray design and deep sequencing have now allowed the creation of replication maps over the entire genome. Several examples of these maps now exist. One approach, performed in the mapping of replication over the mouse genome (Farkash-Amar et al, 2008), has been the segregation of a dividing cell population into temporal segments, followed by BrdU labelling to identify the areas actively
replicating in each temporal fraction of the population. The labelled DNA is extracted, undergoes semi-quantitative PCR and the areas of the genome selected from each population fraction are detected via microarrays. This data can then be used to create a temporal map of replication.

A second approach has been demonstrated by Woodfine et al (2004) to map chromosome 22 and later by Desprat et al (2009) to map the genomes of a variety of human cell lines. Cell populations are fractioned by DNA content (given an S phase population lies between 1n and 2n) into G1 and S phase populations. Each fraction undergoes DNA extraction, with the DNA essentially being sheared, differentially labelled and hybridised to microarrays. Areas of the genome replicated towards the beginning of S phase will be present in higher copy number. The ratio of signal of S to G1 DNA for each probe therefore gives an indication of when in S phase the hybridised DNA replicates.

The sensitivity of each method may still require refinement however. Desprat et al, for example, indicate that the maximum ratio of signal of S to G1 DNA should theoretically be 2. However, their maximum ratio was only 1.6 due to contamination between fractions during the sorting of cells into G1/S. The data is also likely to provide varied interpretation depending on the window size used to bin the output of the microarray experiments. The variation seen between cells in population-wide studies also leads to the question of how synchronised replication is between individual cells of the same lineage.

### 2.7 S Phase Checkpoints

The replication of the entire mammalian genome is a huge undertaking, with many hazards and potential reasons for stalling. As such, a system of steps must be used to ensure that different subunits of the genome have been completed successfully. As noted by Hyrien et al (2003) the timely replication of the entire genome could not be ensured through the use of
one single checkpoint shortly before mitosis. Multiple checkpoints are therefore required, each indicating that a separate sub-phase has been completed.

The exact number and nature of these internal-S Phase checkpoints is still in question however. The existence of at least one has been demonstrated. Shechter et al (2004, DNA repair) suggest a G1/S border checkpoint involving the inhibition of origin firing through Cdc7/Dbf4, followed by two internal checkpoints involving Cdc7/Dbf4 and Cdk2/Cyclin 5 and finally the Cdc2/Cyclin B at the G2/M border.

The initial G1/S border checkpoint theorised by Shechter et al (2004, DNA repair) could also function through a pathway suggested by Bailis & Forsburg (2004) in which Mcm7 is usually sequestered by the tumour suppressor protein Retinoblastoma (Rb), thus inhibiting DNA replication at the licensing step. Cdk4 can activate licensing through phosphorylation of Rb, which releases Mcm7 and allows it to assemble into the pre-RC. Cdk4 would therefore be able to regulate licensing in addition to Cdc7/Dbf4 being able to regulate origin firing. However, the control generated by this mechanism would be mitigated by the occurrence of some licensing during late-telophase (Dimitrova et al, 1999).

The Cdc7/Dbf4 intra-S phase checkpoint is also corroborated by Jares et al (2000), who observe Cdc7/Dbf4 susceptibility to a number of states of replicative stress. Such a checkpoint could operate by the phosphorylation of Cdc7/Dbf4, reducing Dbf4 association with chromatin, thus inhibiting phosphorylation of Mcm2 and preventing further origin firing (Bailis & Forsburg, 2004).

Seiler et al (2007) show how a mid-S Phase checkpoint can be evoked in response to DSBs. The checkpoint is chk1 mediated as inhibiting chk1 allows it to be bypassed. Such a checkpoint could be in response to the activation of the ATM pathway or potentially to the...
ATR pathway if the DSBs are degraded to produce ssDNA. Alternatively, chk1 may be the effector of another completely separate regulation pathway.

Another candidate as a potential mid-S phase checkpoint is the “3C Pause” postulated by Drouin et al (1990) which involves a clear break between replicating R and G bands. The potential for such a checkpoint can be seen by blocking cells with methotrexate or thymidine, then releasing the cells and adding BrdU. The BrdU is only incorporated into G bands, possibly because the cells were stalled at a checkpoint at the R/G transition. Goldman et al (1984) theorised that this “3C Pause” is caused by factories dissociating from early-replicating regions and moving to late-replicating regions. However, this relies on the factories moving in synchronised fashion so as to give a clean break in replication.

The last potential checkpoint for a cell replicating its DNA exists at the G2/M border. Whilst it was initially thought that this checkpoint was mediated by Cdc25c, it has also been found that Cdc25a can also influence the checkpoint (Niida & Nakanishi, 2005), in addition to its roles in the ATM and ATR pathways. Cdc25a and Cdc25c both act through a Cdc2/Cyclin B mediated pathway in order to trigger the checkpoint. However, the checkpoint is also may also be inducible through exposure to IR or UV light via through the Mitogen activated phosphor (MAP) kinases p38g and p38a respectively (Niida & Nakanishi, 2005).

Overall, despite there being a number of postulated checkpoints, there is some question as to the possibility and extent of interlinking between these pathways. Some pathways may merge into a common effector and hence evoke the same checkpoint but in response to a different stimulus.
2.8 Previous modelling relevant to the project

The aims of this project require the integration of two existing areas of modelling as part of the analytical process. Models currently exist that describe DNA replication to some degree, although often these concern lower organisms and lack much of the detail required to understand the mammalian system. To incorporate relevant data, these models will have to be built upon so that a more appropriate format is developed to describe mammalian cells.

Additionally, this project aims to study DNA replication across the nucleus, incorporating a spatial aspect rather than a simple linear model. This will require modelling to simulate the nucleus itself. Models that describe this process will also therefore be analysed in this section.

2.8.1 Modelling of DNA Replication

An increasing number of models have recently been created to describe various aspects of the S phase or DNA replication in general. These are often simulations of lower organisms, but there is potential for common modelling traits to be shared, just as there will be common traits between organism’s replication as an evolutionary consequence.

This review of modelling within the field will therefore begin with a stochastic approximation of DNA replication in *X. laevis*. This model was developed by Herrick *et al* (2002) and features DNA replication along a string of replicons operating through the use of “Kolmogorov-Johnson-Mehl-Avrami” (KJMA) dynamics. The KJMA model is used in the prediction of crystalline growth, operating through nucleation at certain points according to a set probability followed by growth in a bubble around the nuclear point. Upon meeting another bubble, the growing edges merge although the remainder of the two bubbles continues to grow.
The use of the KJMA model was facilitated by the homology between crystal growth and a basic interpretation of DNA replication. In Herrick et al.'s model, DNA was divided into a series of replicons that could each only be replicated once. A nucleation event represented the firing of an origin within a replicon which could occur randomly after each time interval with each replicon having the same probability of firing. Once an origin was formed, replication forks moved along the DNA bi-directionally. The forks continue to propagate until they merge with an oncoming fork at which point growth at the two forks is terminated. The sister forks of the two merged forks continue to grow.

Whilst this model does give a fairly accurate reproduction of experimental data, the nucleation factor was derived from the data that was to be simulated and thus is set according to the results that were observed. A number of different nucleation factors were used to simulate replication within different organisms, such as a high initial value to represent efficient early firing in synchronisation or a low initial value followed by a much higher value in order to fill in any unreplicated gaps. A consistent moderate value does however lead to the formation of gaps due to the stochastic nature of origin firing.

A similar model was demonstrated by Rhind (2006) with a DNA strand divided into 100 origins each with a 3% probability of firing per 5 minute time increment. Each origin represents an undisclosed length of DNA. Growth simply occurred at the rate of one neighbouring origin per 5 minutes. Given these setting, DNA replication does however often suffer from random gaps that leave stretches of DNA unreplicated as the relatively small firing efficiency prevents rescue of these areas by local potential origins. Similar problems were encountered during modelling of the S. pombe genome (Lygeros et al., 2008) using a model populated with experimentally determined and bioinformatically predicted origins. The modelers concluded that either replication continued into G2 or that another mechanism existed to close the gaps in replication that caused the extended S phase.
A solution was suggested by Rhind in the form of a second model termed the “Increased Efficiency Model” (IEF). This model features increased firing efficiencies after predefined length of time in order to close replication gaps. In vivo terms, such a process could be representative of having a constant number of replication forks due to the constant populations of polymerase or due to ATM or ATR regulation. Alternatively such a process could be governed through a diffusible activator such as Cdk2/Cyclin E or Cdc7/Dbf4.

However, the IEM still resulted in a random pattern of replication, which is not applicable to DNA replication at a higher scale as it ignores differentiation between early and late replicating regions. Rhind therefore created another variation of the model termed the “Relative Efficiency Model” (REM). This model featured predefined areas of elevated firing efficiencies. This could be in response to any of the features discussed in Section 6.2 such as chromatin structure or transcription. However, the later replicating regions also need an increased firing efficiency after a time threshold has been reached, as otherwise the regions would frequently remain unreplicated. This model generally produces complete replication of early regions before later regions; however, unless late regions are reduced to ‘no firing prior to specified time point’ there is always the chance of a supposedly late origin firing early.

On a smaller scale, the REM could also be used in response to certain areas being in or near replication factories; if a nearby origin has fired then it will currently lie within a replication factory along with a number of its neighbours. These origins are therefore also more likely to fire (where DNA persistence lengths and the removal of origin interference permit).

Contrary to the idea of increase firing efficiency towards the end of S phase, a model of the early embryos of Xenopus laevis (Goldar et al, 2008) found a best fit to experimental data through a decrease in I(t) towards the end of S phase. The mechanism involved the accumulation of a limiting factor throughout S phase which would promote replication, such
as Cdc45 which is a stable component of replication forks. This factor would also promote an increase in I(t) (contrary to common conceptions of forks and origins acting antagonistically). The positive feedback would continue until around 18 minutes into the 26 minute S phase (duration set according to their experimental data), at which point the rate of fork initiation became equal to their rate of termination. This caused a continuous decrease in I(t) until S phase was completed. Such a mechanism may not be suitable for the simulation of a mammalian system however, as the positive feedback of forks and origins firing would interfere with the possibility of negative feedback loops of origin interference (likely via the ATR pathway) used to balance fork rates and origin densities. However, the possibility of an overall positive feedback loop for replication activity or factory formation could be considered.

Whilst altered firing efficiencies can give a more reliable S phase completion, an alternative has recently been explored by Blow & Ge (2009) on an intra-replicon cluster scale model. Using replicon clusters drawn from the distributions of Jackson & Pombo (1998) they simulate a 250 kb replicon cluster and study the effects of small numbers of efficient origins compared to a larger number of inefficient origins. The number of licensed origins and their initiation probabilities were varied, with inactivated origins passively replicated by elongating forks. From these simulations, sets of potential origin numbers and firing efficiencies were created that resulted in an average of 5 origins firing per replicon cluster, as would be expected on average in a 250 kb cluster. To reduce the parameter space further, they then used these combinations in a second set of simulations whereby the fork rate was reduced to simulate the addition of hydroxyurea. Comparison to experimental data then showed that a set of 10-20 inefficient origins was acceptable to both the control and the hydroxyurea dataset.

From a modelling perspective, this experimental method thus indicates that a single set of origin efficiencies cannot be ruled out during the development of models. Modelling of replication within clusters may therefore need to take into account the potential for common
firing probabilities or for differential ones. Differential probabilities may also occur on a gradient, rather than the simplistic view of primary and secondary origins. The potential for interference between origins via the ATR pathway (see Section 2.3.6) may also factor into these probabilities. In addition, there is also the question of the density of potential origins, as it is still unknown how pre-replicative complexes and observed origins densities are correlated at a single cell level (see Section 2.2.1). Whilst the final results of origin firing are well documented, it may prove a difficult modelling question to infer the original distribution of potential origins without single cell data.

2.8.2 Modelling of chromosome structure, territories and DNA.

Whilst there is no definite model that describes the coordinates of DNA foci and chromosome territories within the nucleus, a number of attempts have been made to generally characterise the distribution of chromatin.

Some of the oldest and perhaps most naive models treated the distribution of DNA within the nucleus as a simple polymer, with its distribution modelled by a random walk. However, this simplistic view could not explain why the displacement between two loci did not necessarily increase with distance between the two loci along the chromatin. Such a model also leads to tangled knots of DNA, which from a replication viewpoint would lead to great complications. There was also no explanation of the formation of chromosome territories, with localised distribution of DNA within a relatively small volume.

A number of models have therefore attempted to build looping into the distribution of chromatin in order to provide a better fit to experimentally observed data. An example is the Random Loop (RL) model proposed by Mateos-Langerak et al (2009). The model was created with two main aims; firstly, to simulate the independence of loci displacement from their
spatial distance along the chromosome, and secondly to simulate correctly the folding of DNA at multiple scales of structure.

Their RL model featured a linear random walk of DNA monomers that each had a probability of interacting with any other non-adjacent monomer (forming a loop). It should be noted that the probability of interaction was randomly determined, which is not entirely biologically accurate, as particular loop formations are more or less likely dependent on the flexibility of the DNA. However, the result of the model at fitted probability settings was the creation of 10-30 loops per 100 Mb, which essentially forced the chromatin into the localised structure of chromosome territories. On the scale of up to 2 Mb, an increase in distance between 2 loci resulted in a greater 3D displacement. However, from 2 to 10 Mb, displacement became independent of the location of the 2 loci on the chromosome. Similar results were found when the chromatin was made to be heterogeneous, with varied probabilities of interactions between loci, and results were verified by 3D-FISH.

An alternative model to those based on random walks (with or without the inclusion of loops) was proposed by Grosberg et al (1993) and has recently been supported by evidence of DNA loci interactions by Lieberman-Aiden et al (2009). Originally termed the ‘Crumpled Globule’ and later the ‘Fractal Globule’, this model sought to avoid the knotting of DNA that occurred with the polymer based models. In order to prohibit these knotted formations, it was proposed that chromatin should be crumpled into a fractal spatial structure. Each level of structure would involve local crumpling of the DNA, leading to a series of structures. These would then themselves be crumpled to give a larger structure and so forth, reminiscent of the properties of fractals. The lowest scale step of crumpling would be defined by the rigidity of the DNA, and the largest scale of crumpling would form the nucleus itself. The fractal globule model not only avoids knotting, but would also facilitate organised unfolding and refolding of chromatin e.g. for gene expression or replication.
This model of chromatin organisation has recently been corroborated by evidence from a high-throughput study of loci interactions across the nucleus performed by Lieberman-Aiden 
et al (2009). Experimentally derived contact probabilities of loci over given distances proved more similar to those predicted by fractal globule model than the RL model. They also noted that the 3D distances predicted between loci were also similar to those observed by 3D-FISH.

Given the supporting evidence for each model, simulations of DNA replication over the mammalian cell nucleus may have to be tested on a number of model structures until a conclusive structure is reached. There is also the possibility of a hybrid of the different structure types on different scales of organisation. This would lead to a number of implications, such as the structure of replicons within clusters and the grouping of these clusters spatially. Varying models of organisation may lead to varying probabilities of factory formation along DNA, especially if factories can activate nearby clusters both spatially as well as next-in-line along the DNA.
3.0 Methods & Materials

This chapter contains the protocols for the experimental procedures used in this project. Each method will in turn be described and its general use explained. Their exact use with regards to gathering data will be specified in further sections as appropriate.

Protocols involving centrifugation steps show a range of speeds from 800-2000 rpm, dependent on the average size of the cell type being used.

3.1 BrdU Labelling

5-bromo-2-deoxyuridine (BrdU) labelling involves the incorporation of bromine labelled deoxyuridine into the actively replicating regions of a cell’s DNA. A pulse of BrdU at excess concentration is added to the cell’s media and then washed out after a specified time period. This will label all the actively replicating regions of DNA within the given period (with varying efficiency). These areas can then be detected through immunostaining as shown in Sections 3.4 and 3.5.

1. Grow cells to ~70% confluence.
2. Add perturbing reagent if required (e.g. caffeine, aphidicolin).
3. Incubate cells with 25 µM BrdU for pulse duration (often 20 minutes) at 37°C.
4. Wash cells with PBS and add fresh media.
3.2 Biotin Labelling

Biotin-11-dUTP (biotin) labelling is a second technique used to label actively replicating areas of DNA. However, rather than using an excessive concentration as for BrdU, cells can be transfected with a limited concentration. When used for DNA fibre analysis, this has the useful property of heavily labelling replication forks at the start of the pulse and then becoming weaker as the biotin pool is exhausted. This characteristic “comet” shape can therefore be used to determine the direction that the replication fork was moving.

Protocol for a 6 well plate:

1. Grow cells to ~70% confluence, aiming for 2.5-5 x 10^5 cells per well required for the experiment.

2. Add perturbing reagents (e.g. caffeine, aphidicolin) as required.

3. Create transfection mix (volumes per well):
   a) Mix 24 µl PBS and 6 µl of FuGENE® Transfection Reagent, incubate on ice for 5 minutes.
   b) Add 2 µl of Biotin-11-dUTP to mix, cover from light and incubate on ice for 10 minutes.

4. 2x Wash cells with cold PBS.

5. Add transfection mix at the centre of the wells.

6. Incubate the 6 well plate levelly on ice for 9 minutes.

7. Wash cells with cold PBS.

8. Add 1-2 ml of cell media¹ and incubate for 30 minutes at 37°C.

9. Wash cells with PBS.

¹ In the case of perturbation (e.g. through caffeine or aphidicolin), add the reagent to the media in equal concentration to step 2.
To recover cells:

10. Recover cells either through trypsinisation or through scraping from well, resulting in the cells being suspended in ~1ml of PBS.

11. Spin down cells for 4 minutes at 800-2000 rpm.

12. Recover cells and dilute in 100 µl of PBS.

### 3.3 EdU Labelling and staining via Click-iT®

A third potential technique for DNA labelling is through the use of 5-ethyl-2'-deoxyuridine (EdU), in this case via the Click-iT® kit (Invitrogen, Paisley, UK). However, due to the use of only a single labelling step, the signal is insufficient for detection by DNA fibre analysis. There is also concern that EdU stalls replication, hence it should not be used for the primary pulse in a dual pulse experiment or in experiments that require continued cell growth.

**Labeling:**

Label as 3.1, using a concentration of 10 µM EdU.

**Preparation:**

1. 3x Wash with PBS.

2. Fix in 4% Paraformaldehyde for 10 minutes.

3. 3x Wash with PBS.

4. Permeabilise cells with 0.5% Triton (100 µl per well) for 5-10 minutes.

5. 3x Wash with PBS.
Staining (24 well plate):

1. 2x Wash with PBS.

2. Mix staining solution (measurements per 200 µl):
   i. 20 µl Click-iT® EdU buffer
   ii. 8 µl Copper Sulphate Solution
   iii. 20 µl Click-iT® EdU Additive
   iv. 2 µl Alexa 488 Fluorophore
   v. dH₂O to 200 µl

3. Add 50 µl per well and incubate for 30 minutes

4. Wash with PBS.

3.4 DNA fibre spreading (Jackson & Pombo, 1998)

DNA fibre analysis is a powerful technique that can be combined with labelling and staining of DNA in order to allow a number of observations to be made. These include the determination of the rate of nucleotide incorporation in actively replicating regions (and hence replication fork speed) and an estimation of origin density.

The fibres are created by spotting and lysing cells at the top of a slide and then tilting to allow the DNA to stretch along the slide. The DNA is dried and fixed, and can then be detected via the subsequent staining steps (see Section 3.5).

1. Store cells at 1000 cells per µl in PBS on ice to slow cell cycle. For smaller cells (e.g. Cho or DT40), use 2000 cells per µl.

2. Spot 2 µl of cell suspension at the top of a slide (suspension can be diluted with unlabelled cells if necessary).

3. Dry for 3 minutes.

4. Add 7 µl of Spreading Solution¹, mix and incubate for 5-7 minutes until the edges of the spot begin to dry.
5. Tip slide to -20° allowing the drop to run down the slide over the course of 3-5 minutes.

6. Dry slide and fix for 10 minutes in 3:1 methanol:ascetic acid. Dry and store at 4°C.

1 Spreading Solution:
200 mM Tris-HCl (pH7.5)
50 mM EDTA
0.5% SDS

3.5 DNA fibre immunostaining

Prior to microscopy, labelled DNA fibres must first be immunostained according to the labelling reagents. DNA fibres labelled with BrdU or Biotin are processed using the protocol shown below.

1. 2x Wash with water for 5 minutes.

2. Wash with 2.5 M HCl. Incubate for 1 hour with 2.5 M HCl (this step is only required if the DNA has been labelled with BrdU).

3. 2x Wash with PBS.

4. 2x Wash with Blocking Solution² for 5 minutes & incubate for 1 hour with Blocking Solution.

5. Mix 1:1000 primary antibody (versus first labelling reagent) in Blocking Solution, using 100 µl per slide. Cover the slides with parafilm and incubate for 1 hour or overnight at 4°C.

In the case of a second label:

a) 3x Wash with PBS
b) Fix for 10 minutes in 0.5 ml of 4% Paraformaldehyde.
6. 3x Wash with PBS.

7. 3x Wash with Blocking Solution for 5 minutes.

8. Mix 1:500 secondary antibody in Blocking Solution, using 100 µl per slide. Cover the slides with parafilm and incubate for 1.5-2 hours.

9. If dual labelling with antibodies, repeat steps 11 to 16 with the antibody versus the second labelling reagent.

10. 2x Wash with PBS.

For DNA staining:
    a) Incubate slides with 1ml of 1:10,000 Yoyo in PBS for 10-20 minutes.
    b) 3x Wash with PBS.

11. Mount cover slips onto slides using 1:1 Glycerol:PBS.

Blocking Solution:
    For 100 ml: 100 µl Tween 20, 1 g BSA, PBS to 100 ml.

3.6 Whole Nuclei Immunostaining

Labelled cells can alternatively have their entire nuclei undergo the immunostaining procedure so as to allow observation of the areas of the nuclei that underwent replication during the labelling pulse. This technique allows the observation of the patterns of DNA replication through S phase.

1. Grow cells to ~70% confluence on cover slips in a 24 well plate, aiming for 2.5-5 x10^4 cells per well required for experiment.

2. Add perturbing reagents (e.g. caffeine, aphidicolin) as required.

3. Perform the first pulse via the appropriate protocol. If performing a dual labelling experiment (e.g. a biotin pulse followed by a bromine pulse), incubate if a delay is required and then perform the second pulse.
4. 2x Wash with PBS.

5. Fix in 4% Paraformaldehyde for 10 minutes.

6. 3x Wash with PBS.

7. 2x Wash with H$_2$O.

8. 2x Wash with 2.5 M HCl, Incubate 1 hour in 2.5 M HCL (only if labelled with BrdU).

9. 3x Wash with PBS.

10. Permeablise cells with 0.5% Triton (100 µl per well) for 5-10 minutes.

11. 3x Wash with PBS.

12. 3x Wash with Blocking Solution for 5 minutes & incubate for 1 hour with Blocking Solution.

13. Mix 1:1000 primary antibody (versus first labelling reagent) in Blocking Solution, using 50 µl per well. Incubate for 1 hour or overnight at 4°C.

14. 3x Wash with PBS.

15. 3x Wash with Blocking Solution for 5 minutes.

16. Mix 1:500 secondary antibody in Blocking Solution, using 50 µl per well. Incubate for 1.5-2 hours.

17. If dual labelling with antibodies, repeat steps 11 to 16 with the antibody versus the second labelling reagent.

18. 2x Wash with PBS.

19. To weakly stain the entire nuclei, incubate with 10 µM Hoechst for 20 minutes.

20. Remove cover slips from wells and mount in 1:1 Glycerol:PBS.
3.7 Serum Starvation Synchronisation

Synchronisation of cells can be achieved through removal of bovine serum albumin from the cells’ media. This technique demonstrated only low levels of synchrony during tests with an MRC5 population however.

1. Grow cells until ~70% confluent.
2. 2x Wash with PBS.
3. Add media without BSA. Incubate for 24 hours at 37°C.
4. Add fresh media with BSA to release.

3.8 Nocodazole Synchronisation

Nocodazole can be used to synchronise cell populations at mitosis through the inhibition of the formation of mitotic spindles. Cells can then be released through washing to remove the nocodazole. Partial synchronisation of MRC5 populations was achieved through the use of this method.

1. Grow cells to ~70% confluence in a 75 ml flask.
2. Incubate cells with 100 ng/ml nocodazole at 37°C (optimum incubation length of 12 hours determined for MRC5 cell line).
3. Gently wash the media from the flask over the adherent cells to remove the cells with a ‘budding’ appearance. Retain media in a centrifuge tube.
4. Pellet cells through centrifuging at 800-2000 rpm for 4 minutes.
5. Remove media, wash cells with PBS.
6. Repeat steps 4 and 5 x2.
7. Pellet cells through centrifuging at 800-2000 rpm for 4 minutes, re-suspend in fresh media and seed as appropriate.
3.9 Aphidicolin Synchronisation

Aphidicolin can be used to synchronise cells at the G1/S phase border through inhibition of replication forks. The most accurate synchronisation of MRC5 populations was achieved through the use of a nocodazole block followed by an aphidicolin block.

1. Grow cells until ~70% confluent.
2. Add 2 µg/ml aphidicolin. Incubate at 37°C.
3. To release cells from block, 2x Wash with PBS and add fresh media.

3.10 Summary of Work Flow

Shown below is a summary of the protocol workflow used for data acquisition.
The outputs of these processes require data analysis through the study of images. The measurement of fork rates and inter-origin distances is possible using existing software (such as the LSM Image Browser Rel. 4.2 (Zeiss, Hertfordshire, UK)) to measure distances by eye. However, the analysis of the replication patterns across entire cell populations was seen to require further scrutiny to determine whether it could be performed by an automated process. The development of this method is documented in Section 6.1, along with an analysis of its feasibility.

3.11 Summary of Modelling Method

The concept of a linearised chromosome is used within the models developed in Chapters 4, 5 and 7, hence this section will describe the process of the creation and parameterisation of a modelled chromosome in detail. Variations on this modelling concept have been published in the paper Shaw et al, 2010. Data to describe the chromosomes was drawn from the UCSC Table Browser (documented in Karolchik et al, 2004) with the March 2006 human genome assembly. The models were implemented in Matlab on a standard desktop PC.

1. The lengths of the chromosome(s) being simulated were downloaded via the UCSC Table Browser.

2. The parameters for the distribution of replicons per cluster and the length of the replicons were drawn from empirical distributions taken from Jackson & Pombo (1998) (see Section 11. Appendix, Table 11.1). The distribution describing the length of replicons was approximated to a normal distribution ($\mu = 140.6238$ kbp, standard deviation = 58.8192) in the original model.

3. The length of the chromosome was used as an upper limit for the generation of DNA replicon clusters. These were created through the following process (which was repeated until the total length matched or exceeded the chromosome length):
a) The number of replicons per cluster was drawn via the generation of a uniform random number (0-1) compared to the cumulative probabilities of Table 11.1.

b) For each replicon within the cluster, a replicon length was sampled from the approximated length distribution. Sampling was performed by generation of a uniform random number.

c) The length of the cluster was summed and assigned to the cluster object, in addition to the number of replicons and the length of each replicon.

4. In some of the models developed, a number of replication factories were created. The exact number was generated by rules specified in each model e.g. the number of factories was 10% of the number of clusters. These were seeded amongst the clusters according to further preset rules (e.g. via a uniform random number-based distribution with reallocation if two factories were placed in the same cluster).

5. With the completed linear structure, time was applied to the model, beginning at t=1 and continuing with 1 minute increments.

6. At each timestep, a subset of clusters were active. This subset was defined for each model by either

   a) *The presence of a replication factory around that cluster.*

   or

   b) *The cluster having activated by beating a probability of activation.* Each cluster would be assigned this probability (according to rules set in the model) and at each timestep, each inactive cluster generated a uniform random number (0-1). If this number was less than or equal to the firing probability, the cluster became active and remained so until all DNA within it was replicated.

7. At each timestep, each active cluster replicated DNA for that minute interval. The distance replicated was calculated via the equation:

   Distance (kbp) = 2 * Fork Rate (kbp/min) * Time interval (min)
The value(s) of the Fork Rates were described in the modelling parameters of each model.

Dependent on the model settings, the distance replicated was either:

a) *Subtracted from the initially longest replicon until its remaining length was zero.*

Under this setting, the total DNA replicated by the cluster during the timestep was approximated to the distance subtracted from the longest replicon multiplied by the number of replicons in that cluster.

If the remaining length of the longest cluster equalled zero, the cluster was marked as complete and inactive. If the model was using the ‘factory’ concept, the factory disassembled.

or

b) *Subtracted from each replicon within the cluster that had any length remaining to replicate.*

Under this setting, the total DNA replicated by the cluster during the timestep was the sum of the DNA replicated within each replicon.

If the remaining length of all clusters equalled zero, the cluster was marked as complete and inactive. If the model was using the ‘factory’ concept, the factory disassembled.

8. At the end of each timestep, the total DNA replicated for each active cluster was summed.

9. When each cluster had completed replication and was inactive, replication was complete and the simulation ended.
It should be noted that in the fork elongation mode of the models in Chapters 5 and 7, replication forks had the potential to progress from one cluster to the next. This used the following mechanism:

1. If a replicon was first or last in a cluster, it was defined as being adjacent to the next cluster (the first replicon was adjacent to the cluster to the left i.e. cluster \( (n-1) \) and the last replicon was adjacent to the cluster to the right i.e. cluster \( (n+1) \)).

2. If either replicon completed replication through the previous method, one of its active forks could progress to an adjacent cluster if the closest replicon had remaining DNA to replicate. E.g. a left moving fork from the first replicon of cluster \( (n) \) required DNA remaining in the last replicon of cluster \( (n-1) \). If this constraint was satisfied, the replicon was marked as undergoing passive replication.

3. During each timestep after being marked, the replicon had its remaining length to replicate reduced by:

\[
\text{Distance (kbp)} = \text{Fork Rate (kbp/min)} \times \text{Time interval (min)}
\]

This amount was also summed to the total DNA replicated for that timestep.

4. If the replicon’s remaining length was reduced to 0, the fork progressed to the next adjacent replicon if the replicon had remaining length to replicate. In this way entire clusters could be replicated passively without activation. If this occurred, the fork could progress to the next adjacent cluster if the constraint in step 2 was satisfied.
4.0 - A simplistic linear model of the Mammalian Cell S phase

4.1 A Simplistic Linear model of Replication

The first step towards developing models to describe and explore the mammalian cell S phase was to develop a conceptual framework for the modelling of the system that would incorporate the basic factors that were initially selected for study. Throughout the project, data was gathered on a number of scales, from descriptors of individual forks through to replication factory dynamics and whole nuclei observations. Thus, the initial framework was set on a chromosome wide scale, allowing incorporation of the smaller scale parameters whilst also having the potential to be scaled up to representation of a whole nucleus. The linear chromosome model concept was developed in collaboration with Robert J Platt and Jennifer K Withers.

In a separate area of modelling, experimental results from Maya-Mendoza et al (2010) have been reproduced in simulations of the nucleus developed in Section 4.2.

4.1.1 Modelling Method

Given the aforementioned parameters that would be explored within the model, a suitable platform for their observation was created through the partitioning of the genome in an imitation of the replicon clusters as observed in vivo. Each replicon cluster was created through the drawing of parameters from empirical distributions (Jackson & Pombo, 1998) which provided the number of replicons per cluster and the length of the replicons. The number of replicons was sampled from the experimentally observed distribution and then an individual sample of length was taken for each replicon to provide a total length for the cluster. Such clusters were repeatedly drawn until the length of each chromosome was
divided into replicon cluster. (See Fig. 4.1a). The data concerning length of replicons was approximated to a normal distribution (μ = 140.6238 kbp, standard deviation = 58.8192) for this process, whilst the number of replicons per cluster was sampled from binned data (see appendix for the distributions used). For a detailed method, see Section 3.11.

It was assumed, as shown experimentally (Jackson & Pombo, 1998, Maya-Mendoza et al, 2007) that all replicons in a given cluster fire at approximately the same time. Hence, the length of time required for a cluster to fully replicate would be the replication time required for the longest replicon within the cluster. This is an approximation that will apply in the majority of cases, though it is noted that this will not always be true as in a minority of cases adjacent replicons display different rates of synthesis (Conti et al, 2007). Published rates for the average speed of fork elongation generally fall in the range 1-2 kbp/min in mammalian cells. For this initial model, an average fork rate of 1.7 kbp/min (Jackson & Pombo, 1998) was applied throughout S phase. Given an average fork speed, and knowing the length of the longest replicon within a cluster, the assumption of co-ordinated firing allows the estimation of the time required for the cluster to replicate once active.

Given this initial framework, the model then required the addition of the functional units of DNA replication. These are the DNA replication factories, which perform DNA replication after assembling at a replicon cluster. It has been estimated that 10 - 15% of replicons are active at any point in S phase (Jackson & Pombo, 1998). For this initial model it was therefore conservatively estimated that the number of factories active in a chromosome at any given time is equal to approximately 10% of the number of clusters in that chromosome. These factories were initially seeded at random across the framework of clusters (see Fig. 4.1b).
The model was then prepared to be operated under a timescale, with factories randomly moving to another unreplicated cluster upon the same chromosome when the allotted time required for replication of a cluster is completed (see Fig. 4.1c). When all clusters had been replicated, the simulated S phase was complete. For this simple model, outputs such as DNA replicated per minute and total completion time for each chromosome were recorded.

The model was written in Matlab and simulations were run on standard laptop and desktop PCs. Averages shown in the results were determined from sets of 1,000 stochastic simulations.
4.1.2 Results

Given that 10% of the genome is actively replicating at any time within the model, and that the average cluster created required around 60 minutes to replicate, on average the completion time for a chromosome was ~10 hours, which is to be expected for a mammalian cell. The rate of completion of clusters over time was also measured (see Fig. 4.2), which, as a result of constant activation of clusters, was linear once the completion of clusters falls out of synchronisation.

Fig. 4.2 – Cumulative number of cluster completed over time for Chromosome 4, averaged over 1,000 simulations (blue line). Red lines indicate a standard deviation either side of the mean.
4.1.3 Conclusions & Perspective

The initially estimated parameters gave an approximation of the mammalian cell S phase comparable to experimental observations. Further details could then be added to the model to attempt to incorporate more complicated factors that are relevant to the system. Even with this relatively simplistic model however, some observations could be made. The completion of clusters over time predicts that a synchronised activation of clusters at the beginning of S phase would lead to a fall in factory activity after an hour at the mean cluster completion time. Although Fig. 4.2 only shows a slight plateau effect after an hour, combined with the time taken to dissemble factories and reassemble at a different location, this would likely lead to a substantial decrease in visible factory activity. Given that a decrease in the number of factories is not observed when cells are studied in vivo, this could imply that the formation of factories at the start of S phase is only loosely temporally coordinated.
4.2 Is cluster activation random?

As discussed in Section 2.6, DNA replication across the nucleus occurs in a series of very specific patterns. These patterns are unlikely to be generated through the random activation of replicon clusters as occurs currently in the simplified model. Experimental data (Maya-Mendoza et al, 2010) also demonstrate the close association of the activation of replicon clusters, with clusters that activate an hour apart being visually adjacent. An average inter-cluster distance of 150 nm was measured from centre point to centre point (n = 161). From a modelling perspective, this prompts the investigation of whether such observations can occur through the random seeding and random activation used in the model in Section 4.1. If such a mechanism does not operate through stochastic activations of clusters, the model must then be adapted to incorporate additional mechanisms. The total possibilities of activation therefore range from a completely random assembly anywhere in the genome through to a carefully ordered placement due to a specific sequence of activation. In order to rule out some of these scenarios, simulations can be used to compare distribution methods to available data.

4.2.1 Modelling Method

The focus of these simulations was to recreate the experiment which allowed the aforementioned observations of association. This involved the creation of a simplified model nucleus, using a similar concept to Section 4.1, with a series of replicon clusters being created from a linearised chromosome. Each cluster of each chromosome was then assigned a spherical volume 120 nm in diameter (Koberna et al, 2005), creating, essentially, a string of beads. These strings were then seeded through an empty nucleus with a diameter of 6μm. Rather than plotting a single chromosome at a time, which may result in globular early chromosomes followed by much more dispersed later chromosomes, all chromosomes were plotted simultaneously, with the first cluster of each being seeded followed by the second etc. Each initial cluster plotted was random, with the sequential plots being made using a random walk. The constraints for this were that the cluster must be in contact with the
previous cluster whilst not overlapping with any cluster by more than 15% (Zink et al., 1999) which allows for the likelihood that clusters are not perfectly spherical and that there may be a degree of intermingling between clusters. See Fig. 4.3 for a summary of this process.

With a simplified model nucleus created, it was possible to begin a simulation of the experiment. The in vivo process involved using two labelling pulses with an hour interval, in order to identify the subset of clusters that were actively replicating at each timepoint. Cells were then grown for 6-7 days to allow chromosome segregation to occur. Cells with nuclei containing labelled chromosomes were then identified and it was found that the labelled replication clusters are closely co-localised. To simulate this process, the model nucleus also underwent two labelling pulses. Each ‘pulse’ labelled 10% of the clusters (as roughly 10% of clusters should be active each hour). Given a human genome of 6 billion base pairs, divided into replicon cluster of ~1 Mb, it can assumed that roughly 600 clusters should therefore be labelled in each pulse.

Fig. 4.3 – Top left - first step of the simulation, with the initial plotting of the first cluster of each chromosome. Top right and bottom left - the subsequent steps of plotting, with clusters being added sequentially.
4.2.2 Results

The simulations shown below sought to study the association between the cluster labelled in the first pulse and those labelled in the second. In the first simulation, the distribution of active clusters from each pulse was entirely random. This represented subsequent clusters being activated in an entirely random fashion with regard to the prior completion of clusters, as was performed in the simple linear model in section 4.1. Fig. 4.4 shows the results of this simulation, with the first pulse being indicated by blue clusters and the second pulse by yellow clusters.

![Cluster distribution](image)

**Fig. 4.4** – Clusters labelled through random activation.

**Top**- The nucleus is shown with three labelled chromosomes (the rest having been removed to represent chromosome segregation).

**Left**- A closer view of typical interactions within a chromosome, showing little association between blue and yellow clusters.
This simulation represented one extreme of the possibilities of factory progression; that previous activation have no effect on subsequent activations other than removing the possibility of that particular cluster firing. In this case, the mean distance between a blue labelled cluster and the nearest yellow cluster was 215.5 nm (with a standard deviation of 113.7 nm). These values were averaged from three chromosomes of foci drawn from 1,000 iterations of the model. The second simulation explored the other extreme, where the only clusters that can be activated in the second round of activations are those on either side of a cluster that was selected during the first. This would represent neighbour activation of clusters, due to, for example, the encroachment of forks from an already active/completed neighbouring cluster. The results of this simulation are shown in Fig. 4.5.

![Graph showing simulated clusters](image)

With associated labelling, the mean distance between a blue labelled cluster and the nearest yellow cluster was reduced to 139.1 nm (with a standard deviation of 79.6 nm) when again measured over three chromosomes drawn from 1,000 iterations of the model. The mean distance observed through the associated labelling method provides a close match to the experimentally observed mean of 150 nm.
4.2.3 Conclusions & Perspective

Of the two extremes tested in these simulations, the neighbour activation model generated results very similar to the experimental data. Given this example and corroborating literature (Manders et al, 1992, Sporbert et al, 2002), subsequent models of mammalian DNA replication must therefore attempt to simulate non-random activation, although not necessarily through the extremity of neighbour-only activation (which, as a sole mechanism for activation, would likely lead to potential gaps in replication).

The next area of modelling therefore discusses a variety of modes of factory dynamics, with factory ‘movement’ (i.e. their disassembly and subsequent reassembly) being limited to some degree. Several different methods were used to model the movement of the replication factories within the boundaries of the chromosome: nearest-neighbour; localised stochastic movement and random walk. Each simulates a different degree of restriction to replication factory movement.

Fig. 4.5 – Clusters labelled through neighbour activation.

Top- The nucleus is shown with three labelled chromosomes (the rest having been removed to represent chromosome segregation).

Left - A closer view of typical interactions within a chromosome, showing close association between blue and yellow clusters.

Bottom Left – Experimental results observed by Maya-Mendoza et al (2010) after sequential pulses with AF488-dUTP (green) and Cy3-dUTP (red).
4.3 How should Replication Factories ‘Move’?

In the previous section, an entirely random activation of replication clusters was deemed an unlikely method by which DNA replication operates. From a modelling perspective, several different potential methods of movement could be considered as simplified cases of what may occur in vivo. However, in a biological context it is also apparent that the dynamics of factories are likely to be a function of various factors concerning the substrate DNA, as discussed in Section 2.6.2. Of these potential factors, the most commonly accepted factor to correlate with replication at the time of study was with the R- and G-banding of chromatin. It is generally accepted in the literature (Strehl et al., 1997, reviewed by Drouin et al., 1994) that R-banded regions of euchromatic, gene rich DNA replicates early, whereas G-banded regions of heterochromatic, gene poor DNA replicates later in S phase.

4.3.1 Modelling Method

The next step in the development of the model involved the integration of the varied methods of simulating factory progression along with the information describing the localised banding of DNA. To this end, the banding type of each cluster was imposed over the framework developed in Section 4.1 (see Fig. 4.6). In order to accomplish this, the UCSC Table Browser with the March 2006 genome assembly (documented in Karolchik et al., 2004) was first used to obtain the Giemsa staining data for each chromosome. Each replicon cluster in the model was assigned one of the following band types after the simulated chromosome was compared to the staining data: R-band; 25% stained G-band; 50% stained G-band; 75% stained G-band; 100% stained G-band; or other (e.g. stalk). The G-band percentages indicate the relative amount of staining present in comparison to other G-bands. At this initial stage, the percentage staining was not used as an input parameter in the modelling, but was included to allow observation of variation in replication between differentially stained G-bands due to their spatial distribution. The assignment of either R- or G-banding to a cluster could then potentially be used to assign preferential firing parameter, simulating the variation seen within the firing probabilities of different types of chromatin.
As the model operates through the recording the completion of clusters over time, the speed at which replication occurs was an important parameter. To allow observations to be as accurate as possible, a variable distribution of fork rates (Takebayashi et al., 2005) was applied to the model. These measurements were taken from HeLa cells and span 7 hours of S phase. Pre-allocated times were removed from clusters and the time required to replicate an active cluster was calculated dynamically, dependent on the variable fork rate that was applicable at the time of the clusters activation. Once drawn, this fork rate was then held constant for that cluster.

The aim of this model was to observe how programmed factory behaviour could lead to experimentally observed phenomena. Three areas of interest in particular that are suited to observation in a linear system are the association between sequentially activated clusters, the type of transition between activation of R- and G-banded areas of the genome and the number of gaps in replication. Ideally, clusters should be in close proximity to a previously activated factory to match the first observational aim. For the second aim, as far as possible, R-banded clusters should be replicated prior to G-banded clusters, leading to a sharp transition between the two which occurs over a short time period. The possibility of gaps will likely depend on the programmed dynamics of factories, although the overall duration of S phase will be extended if limited factory mobility inhibits the activation of clusters.

<table>
<thead>
<tr>
<th>Cluster: 1</th>
<th>Cluster: 2</th>
<th>Cluster: 3</th>
<th>Cluster: 4</th>
<th>Cluster: 5</th>
<th>Etc.</th>
</tr>
</thead>
</table>

![Fig. 4.6 – The division of the chromosome into replicon clusters with the addition of the R- and G-banded template.](image-url)
Given these aims and the extra layers of information added to the model, the dynamics of factory seeding and subsequent progression could now be further investigated. The initial distribution of factories was performed through one of two methods:

i) Factories are placed randomly within R-bands; or

ii) Factories are placed on the chromosome in a completely random manner.

Three different methods of factory progression were also tested. These were:

i) **Nearest Neighbour (Fig. 4.7 i)** - Factories reassemble at the nearest unreplicated cluster (with a random chance of going to either in the case of equidistant nearest clusters). The model could also be informed to progress only to the nearest R-banded cluster if a preference had been given towards R-bands (e.g. in the early period of S phase). Such dynamics would represent a flawless progression of replication factories along the DNA in a linear fashion, always locating the nearest suitable cluster.

ii) **Localised Stochastic Movement (Fig. 4.7 ii)** - This approach attempted to accommodate the possibility of dissociation of replication factories from DNA and reattachment at more distant locations. This was represented by sampling movement distances from a Gaussian distribution, with a mean centred at the current factory location. The standard deviation of the Gaussian was varied between 5 clusters, which ensured that the likelihood for near neighbour activation would be high, and 100 clusters, to allow a wider range of jumping possibilities. If the drawn cluster had already been replicated, another cluster was drawn from the distribution. Additionally, in the case of banding preferences, a probability of firing could be set for clusters. As stated earlier, experimental evidence suggests that R- and G-bands are replicated at different stages of the S phase, hence different probabilities could be assigned to each. In the event of a failure to fire, another cluster was again drawn. The relatively static nature of this method given a narrow distribution simulates the problems faced by the cell if there is insufficient attraction of factories by distance clusters, leading to gaps in replication.
iii) **Random Walk (Fig. 4.7 iii)** - Factories reassemble at a cluster drawn from a Gaussian distribution centred on their current location, as for the localised stochastic movement. However, in the case of a failure to fire (either due to the cluster already having been replicated or a failure to fire due to banding probabilities), the Gaussian distribution re-centred on the selected cluster prior to the next draw. This method simulates the potential for scanning of DNA for the preferred next target and mitigates the potential for gaps in replication. The random walk method was also used with a varying standard deviation between 5 and 50 clusters.

These three potential methods of explaining factory dynamics are summarised in Fig. 4.7.

Fig. 4.7 – Potential methods to simulate behaviour of replication factories
The model was written in Matlab and simulations were run on standard laptop and desktop PCs. Averages shown in the results were determined from sets of 1,000 stochastic simulations.

4.3.2 Results

The model was tested under a number of experimental conditions in order to answer a series of biologically relevant questions:

i) Does the model provide accurate values for overall S Phase duration?
Throughout this study, total replication times for each chromosome were averaged to between 9.5 and 9.8 hours when using the averaged replication fork rate. This is consistent with the previous model, as the total number of factories active at any time had not been altered; merely their distribution and mode of target selection had changed.

ii) Can Factory Guidance by R/G-banding give realistic replication patterns?
One of the key features of mammalian DNA replication that the model aimed to simulate was the transition between R-band and G-band replication. Although there are varying accounts of this process, it is generally observed that the transition of replication activity between R- and G-bands occurs around 4 hours into S phase. The closely synchronised nature of the transition necessitates that the duration of the transition should be as short as possible. The simulations were therefore evaluated with regard to the beginning of the transition, its duration and its midpoint (when the majority of replication factories have left R-Bands). Several methods could be used to generate this feature with respect to the underlying biology. One possibility is that the creation of replication factories is directed towards transcription sites at the G1/S transition so that the majority of factories begin in R-bands (Hassan et al, 1994). However, the model suggests that while this could be a contributing factor, the seeding of factories in R-bands alone it is not sufficient to provide a sharp R/G-band replication transition (see Fig. 4.8a). In this simulation, replication factories were seeded in R-bands and then allowed to progress randomly through nearest neighbour dynamics with no further preference towards R-
or G-band clusters. Nearest neighbour dynamics were chosen as they provide the highest possibility for factories to remain in R band clusters, thus maximising the potential for R-band seeding to concentrate factories in R-band clusters until all had been completed. However, even with these favourable settings, factories rapidly progressed to activating G-band clusters, implying that the seeding alone is insufficient and further guidance of the factories is required. Fig 4.8b shows demonstrates the relatively low level of variation between simulations. The variation upon completion is due to slight differences in the number of clusters required for each simulation (given the random generation of lengths), and as the standard deviation at this point is generally within 1-2 fold of that of the rest of the timecourse, the majority of variation can be accounted to different numbers of clusters rather than to the replication dynamics.

Fig. 4.8a – The average number of replicon clusters replicated over time for chromosome 18. Blue curve represents replication of R-banded clusters. Green curves are for G-bands (light green to dark green indicating 25%, 50%, 75% and 100% Giemsa staining). Red curve represents replication of other regions e.g. centromeres, stalks and variable regions. The time course shown is averaged for 1000 simulations of the model.

Fig. 4.8b – The standard deviations of the replication curves generated for Fig 4.8a. Key as above.
What level of banding preference is required for more realistic dynamics?

Given the requirements for some level of guidance, further exploration of the model first required a study of the level of preference of R-band clusters over G-band cluster required to give more realistic dynamics. For the nearest neighbour mechanism, a complete preference was initially set as it was perceived that a mechanism that simulated direct neighbour activation should not be as unreliable as the two stochastic methods. To determine a biologically relevant threshold for the preference for R-bands over G-bands using these mechanisms, the model was tested using a range of different probabilities for G-band firing. These ranged from an equal preference for G-band firing to that of R-bands through to a 150-fold preference for R-bands over G-bands. Factory dynamics were dictated by the Localised Stochastic Movement model which was used to give a baseline of the factory behaviour under the various firing probabilities prior to the loosening of movement restrictions through re-centring. Fig. 4.9 shows the observed beginning of the transition (the firing of the first G-band), the end of the transition (the firing of the last R-band) and the midpoint between the two over the tested range of firing probabilities for chromosome 7.

Fig. 4.9 - The effect of varied G Band firing probabilities on the R/G band transition. Observed replication landmarks for chromosome 7 under varying G Band firing probabilities. The beginning, midpoint and end of the transition are respectively defined as: the time at which a factory first moves into a G band cluster (stars); the time 50% of factories have moved into G band clusters (circles); the time at which all factories have moved out of R band clusters (crosses). Squares indicate the total time required to replicate the chromosome. Values plotted are averaged from 1000 simulations.
Despite the varying firing probabilities, the midpoint of the transition was seen to vary little, whilst the length of the transition increases as the G-band preference approaches that of R-bands. To simulate the observed transition period between R- and G-bands, a strong preference of R-bands over G-bands was selected with R-bands being 100 fold more likely to be activated than G-bands. This value was chosen as a compromise between minimising the transition duration whilst still allowing a degree of stochasticity.

As a result of this segment of study, upon the application of the R and G band template to the clustered DNA, preferential firing of R-band clusters was set either as absolute (in the case of the nearest-neighbour) or a strong (factories in R-bands 100 times more likely to fire than those in G-bands).

**iv) Does the Nearest-Neighbour Model give realistic factory behaviour?**

The model was then used to explore if nearest-neighbour factory dynamics could simulate temporal changes in patterns of replication sites. Given the total preference for R-bands under this progression setting, this mechanism effectively became a simulation of the scanning of the DNA in a linear fashion, seeking out R-band clusters. When all R-band clusters were completed, a switching mechanism would then allow the replication of G-band clusters, with factories progressing from their last location but now with an unrestricted scan. Simulations for all chromosomes except Chromosome Y resulted in the R/G transition commencing between 3 and 5.5 hours (see Fig. 4.10), roughly corresponding to observations made by Drouin and colleagues (Drouin *et al.*, 1990). Chromosome Y was seen to begin its transition after only 2.7 hours, owing to its high relative content of G-bands. However, the results did show significant variation in the duration of the transition phase between chromosomes. Between chromosomes (with the exception of the Y chromosome), the mid-point transition times (the time when 50% of factories have moved to G-bands) had a standard deviation of 25 minutes, with mid-point transition times ranging from 3.4 hours for chromosome 14 to 5.1 hours for chromosome 17. It should be noted that this was larger than the standard deviation in the overall replication times, which was 4 minutes. The Y
chromosome proved to be a major outlier (see above) and including the Y chromosome in the analysis resulted in a larger standard deviation of 29 minutes.

Fig. 4.10 - The dependency of transition time on the proportion of R-band. For each chromosome, the beginning, midpoint and end of the transition are respectively defined as: the time at which a factory first moves into a G-band cluster (stars); the time 50% of factories have moved into G-band clusters (circles); the time at which all factories have moved out of R-band clusters (crosses). Results are averaged from 1000 simulations.

The differing times of transition initiation between chromosomes can be directly attributed to the relative contents of R- and G-bands. This correlation is trivial, as the time span of R-band replication must depend on the size and number of R-bands. However, this does raise interesting questions concerning the synchronisation with which chromosomes transit from early to mid/late S phase. Any timing discrepancies could be avoided, however, if factories are able to move between chromosomes. The length of the transition is proportional to the overall length of the chromosome, with longer chromosomes having more factories undergoing transition and hence having a higher chance of containing unusually long clusters.
This strictly controlled mechanism of factory dynamics also has the advantage of giving high degrees of neighbour activation due to the domino-like activation that occurs. Without any preference between cluster banding, 98.6% of clusters were activated either when their neighbour was active or within an hour of its completion (averaged over a thousand simulations of chromosome 4). When the total R-band preference is applied, this falls only to 88.4%, due to the requirement of some more distant scanning to locate G-band clusters after the transition switch is toggled and the jumping of G-Band clusters in the interim.

From a biological perspective, the nearest neighbour mechanism gives results that would match observed phenomena. Due to the strictness by which it operates, there is a high degree of associated activation and a sharp R/G transition. However, the mechanism of the total-preference R-Band neighbour activation is unlikely as its reliance on scanning would require at least partial factories to remain and progress along DNA strands, whilst evidence points to their disassembly (Leonhardt et al., 2000). An alternative would be for factories to form around elongating forks, but this may detract from the sharpness of the R/G transition if R-band clusters are situated next to G-Band clusters which would then be sequentially activated. Whilst likely to be less efficient, the stochastic methods may therefore be more mechanistically realistic.

v) Do stochastic models of factory movement give realistic behaviour?

The two stochastic approaches, the random walk model and the localised stochastic movement model, were then applied to factory movement. Interestingly, the localised stochastic movement model could fail entirely if all R-band clusters must fire before G-band replication is permitted. The reason for this failure is that a factory may need to ‘jump’ over G-bands to get to R-band clusters. For Gaussian distributions with a small variance, the probability of a successful jump over groups of intervening G-band clusters may simply be so low that the model becomes computationally intractable. Biologically this would represent the absence of sufficiently strong guidance for factory progression, resulting in factories being incapable of locating distant or remote unreplicated clusters.
Using a random walk to reset the mean of the jumping distribution removed this potential error. Factories were essentially made more mobile, thus mitigating the effect of a narrow Gaussian distribution. Use of this model with narrow distribution (for example a standard deviation of 1-5 clusters) is most like a traditional random walk in which the factory can ‘scan’ along the chromosome, with equal probability in each direction. This scanning occurs by a relatively stochastic method however, compared to neighbour activation, with factories sampling the nearby clusters. A wide distribution (e.g. a standard deviation of 50 clusters) effectively causes the factories to move larger distances between bands. With a narrow distribution, the factory may need to scan through a G-band to reach an un-replicated R-band, but this is not required with a wide distribution. If the difference between R- and G-band replication is caused by a variation in the efficiency of interactions (e.g. chromatin being more condensed and so origins are less accessible in G-bands) then there is only a small probability of interaction between a G-band cluster and a factory scanning through the G-band. Even with a G-band interaction probability of 1%, this has substantial effects (see Fig. 4.11a) and resulted in a smoothing of the R/G transition. If factories do not need to scan through G-bands no smoothing effect was seen (Fig. 4.12a).

Fig. 4.11a - Replication curve for low standard deviation (5 clusters) random walk behaviour over chromosome 2.

Blue curve represents replication of R-band clusters. Green curves are for G-band clusters (light green – dark green indicating 25%, 50%, 75% and 100% Giemsa staining). Red curve represents replication of other regions e.g. centromeres, stalks and variable regions. The time course shown is averaged for 1000 simulations of the model.
Fig. 4.11b – The standard deviations of the replication curves generated for Fig 4.11a.
Key as Fig 4.11a.

Fig. 4.12a - Replication curves for high standard deviation (50 clusters) random walk behaviour over chromosome 2.
Key as Fig 4.11a.

Fig. 4.12b – The standard deviations of the replication curves generated for Fig 4.12a.
Key as Fig 4.11a.
The two stochastic methods produce very different factory dynamics with regard to the association of sequential activation and the sharpness of the R/G transition. The Localised Stochastic Movement model proved unreliable unless factories had the ability to jump large groups of clusters. The smaller chromosomes in particular, with fewer factories, were susceptible to the bunching of factories at one end, which could prevent the replication of distant clusters. Reliable replication was observed with a standard deviation of 100 clusters, which led to a rapidly completed transition period (as all R-bands were easily accessible), but a relatively low level of associated activation (only 50.8% of clusters by the aforementioned criteria).

The Random Walk model was also tested over a range of distributions. At a low standard deviation of 5 clusters, associated activation was higher at 63.2% but due to the inability to jump over intervening G-band clusters, the transition period was much wider, with all R-band clusters being completed only after 8.19 hours on average. With a standard deviation of 10 clusters, associated activation fell to 57.2% whilst the transition was also reduced, with R-band cluster being completed after 6.67 hours on average. By a standard deviation of 50, these valued had approached those of the Localised Stochastic Movement model, given the extra freedom of movement allowed to the factories. The use of a large standard deviation parameter also led to less variation in the types of clusters replicated over time (see Figs 4.11b and 4.12b), as the factories were capable of making more consistent choices of R-banded clusters over G-banded clusters.

vi) **How many replication factories are required?**

In the previous models, the number of factories operating on a chromosome had always been proportional to the number of clusters, given the assumption that 10-15% of DNA is actively replicating at any one time. However, estimates of the number of active replication factories have been attempted *in vivo* by counting replication sites in S phase cells. These estimates vary widely across different cell lines, with active replication sites at a single time-point being estimated to ~150 in a chick cell nucleus (Cossman *et al*, 2000) to over 1,000 sites.
during early S phase of a mouse 3T3 fibroblast (Ma et al., 1998). A mathematical estimate would lie between these figures: assuming an average fork speed of 1.7 kb/min and an average of 4 origins per replication cluster (each with two forks) (Jackson & Pombo, 1998), an average cluster will replicate 8.16*10^5 base pairs of DNA in an hour. Given a haploid human genome of approximately 3 billion base pairs and a 10 hour replication time, this would then require around 370 factories/haploid chromosome set on average throughout the 10 hours (assuming constant factory size). A diploid somatic cell would therefore need to maintain ~740 factories to sustain synthesis across its genome. These values can be compared via the model to estimate the required replication times (see Fig. 4.13), which are independent of the mechanism of factory movement used or band preferences.

Fig. 4.13 - Replication times corresponding to varied factory numbers. Time required to completely replicate chromosome 18 (stars) with estimated line of best fit (dash-dotted line). Crosses indicate approximation of S phase transition (the time point where the majority of factories have left the R Band clusters) with estimated line of best fit (dotted line). Results are averaged over 1,000 simulations.
The results indicate that for S phase to be completed in 9-10 hours, a single copy of chromosome 18 (1.27% of the diploid genome) would require between 20 and 22 factories, which equates to between 1575 and 1732 factories to replicate the entire genome. With a high preference towards R-band synthesis (100:1) this condition corresponds to an R/G transition at 4-4.2 hours. This number of factories is approximately double the value obtained from the mathematical estimate, as a result of the low replication rate during early S phase when using the variable fork rates and the lack of optimisation caused from the random clustering of replicons. In some case, short replicons that would replicate quickly have been grouped in clusters that contain larger replicons, leading to an increase in replication time for the entire cluster. This may represent a limitation of current models; however, as detailed information concerning the size variation of replicons within clusters is at present unavailable.

vii) Can GC content and Gene density approximate R/G-band factory dynamics?

In addition to the guidance of factories via the R- and G-banding, the model was also tested with factories guided by the GC and gene content within replicon clusters. Using these two variables in turn, unreplicated clusters with the highest content of the guiding factor was selected for replication; this representing the average behaviour of a factory in relation to an attractant (the guiding factor). Factories were initially distributed in clusters with the highest guiding factor. The results of these varied methods of factory guidance across chromosome 2 are shown in Fig. 4.14, with a nearest neighbour (hence R-band total preference) trial shown for comparison. The GC and gene density guided factories are therefore scored against their replication of R- and G-band clusters.
Both GC and gene guidance provide a slight trend towards early R-band cluster replication. However, chromosome specific features, such as a large area of gene poor clusters that occupy an R-band rich area of chromosome 2, lead to variation in the replication curves. It is also notable across all chromosomes that the gene guidance scheme results in a lower overall number of clusters replicating during early replication. This is due to the greater possibility of small clusters occupying an area of non-coding DNA. However, in larger clusters gene density is likely to average out and for GC content the preferential factor is less likely to occur in localised areas of high density. Such correlations raises the question: would a small cluster with a high density of preferential factor or with an easily accessible structure be more likely to attract factories than a larger cluster of equal absolute content or overall accessibility?

Fig. 4.14 - Guidance of factories over chromosome 2 by banding and by GC and gene content. Curves represent replication times averaged over 1,000 simulations. Blue lines represent guidance by R-band, red by GC content and green by gene content. The solid lines of each colour represent replication of R-band clusters for that factor, with the dashed lines representing the sum of the differentially stained G-band clusters.
4.3.3 Conclusions

A number of conclusions can be drawn from this simplified model of the mammalian S phase which are likely to be relevant to further modelling and experimental work.

i) The model is capable of generating an R/G transition at ~4 hours and also correctly estimates the overall length of S phase to fall within the experimentally observed range between 9.5 and 9.8 hours, which is typically seen in somatic (diploid) human cells.

ii) It has been shown that the seeding of factories in R-band clusters is alone insufficient to simulate observed replication patterns and that a strong preference for R-bands in also required. Early modelling results indicated that merely beginning with factories situated in R-bands is not sufficient to simulate the observed behaviour, as random factory movement causes them to gradually diffuse out of R-bands. Thus, the model corroborates evidence that during early S phase the assembly of replication factories is biased towards R-bands or a specific feature of replicons that are enriched in R-bands.

iii) Stochastic methods of factory movement, which simulate factories that scan the genome in a multi-dimensional space, allow increased factory mobility whilst also allowing for observations of factory disassembly. However, the added freedom can lead to an unrealistic level of de novo activation of clusters and a smoothing of the R/G transition. The Nearest neighbour model is also limited, as it forces some factories through G-band clusters early in S phase unless a total preference is set for R-bands, yet it gives more realistic neighbour activation. It is therefore likely that a hybrid of the two methods is employed, with the majority of activations occurring through nearest neighbour activations whilst some stochastic activation occurs in reaching more distant clusters.

iv) The number of factories required to replicate a diploid human genome is estimated to be approximately ~1,000 at any single time point, although the structure of replicons clusters can greatly influence this figure. The observed total may also be affected by the altered morphology of factories during S phase, with the possibility of factory aggregation influencing
the experimental data (Hozak et al, 1994). With these conditions, the model also predicts an average number of active factories that falls within the range seen experimentally.

v) Finally, it is shown that gene density and GC content of DNA can be used to simulate the replication programme by guiding factory assembly according to the location of chromosomal R- and G-bands. As all these factors are related and all can be approximated to DNA replication, this may therefore be an issue for consideration when programming the guidance of factories in future models.

The model has a number of limitations however that must be highlighted for further consideration through this study.

i) Whilst the mechanisms discussed could provide realistic dynamics when compared to experimental observations, a number of traits concerning the factories are overly simplistic when applied to a model system. The use of an absolute number of factories (whether programmed or calculated from a percentage of total activity required) creates a system that, whilst understandably optimised, is flawless in its execution. The fact that factories are programmed to always be active, combined with their infallible ability to find another available cluster (although sometimes computationally difficult) leads to an unrealistic plateau of constant replication. This consistency is reinforced by spontaneous assembly and disassembly of factories, and the spontaneous diffusion of factory components to the next target cluster.

The overall effect of these mechanistic issues combined is to create an overly reliable system where factories cannot remain dormant even if a target cluster is very distant. Whilst DNA replication is likely to be a highly optimised process, such flawlessness seems highly unlikely. Even if targeting is not a problem, the time required for formation of factories around clusters would still lead to variation in overall replication levels.

ii) Whilst the model was tested on all chromosomes, each set of simulations was run separately and judged according to the same criteria. However, some chromosomes follow different replication dynamics (such as chromosomes containing large amounts of G bands
generally replicating later in S phase). The separation of the chromosomes during simulations also prevents the progression of factories from one chromosome to another, which may be a potential mechanism through which uniform transition and endpoints could be reached. Parameterising such a system would likely require the development of a carefully structured three dimensional environment however. Whilst the current stochastic mechanisms can represent the progression of factories once the linear DNA is contorted, jumping to another chromosome would require approximated likelihoods of movement to other points in the genome.

iii) The significance of the width of the distributions employed for the stochastic simulations is not sufficiently explored by the model, as the parameterisation is limited by the lack of experimental data. The width of the distribution represents the probability of a factory moving over various distances. Given that DNA is in fact contorted, the distribution need not necessarily decrease with distance, and may have local peaks where clusters are brought into close proximity due to the larger scale structure of DNA, into chromosome territories for example. It could be assumed that for every possible factory location, the distribution would therefore be a function of the localised structure of the DNA (amongst other factors).

iv) Definition of the preference of R-band clusters over G-band clusters is also limited by the lack of suitable experimental data. If there is a non-zero probability that factories will attach to G-bands, then there will be significant smoothing of the R/G transition unless factories tends to move large distances and ‘jump’ past G-bands. This might again only be biologically realistic if the three-dimensional arrangement of the chromosome in the nucleus is considered. Without the ability to diffuse over longer distances, factories would instead have to become dormant until the transition period if a defined switch between R- and G-band clusters is to occur.

v) In order to judge the requirement for fine tuning of firing probabilities and the wide of the stochastic distributions, there is need of a more accurate description of the R/G transition. Using different firing probabilities for pre-RCs within each class of band allows
creation of a fairly sharp transition, but it is uncertain how well this reflects the situation in individual cells. Notably, the present literature on this point is unclear, with some studies describing a distinct ‘3C-pause’ in the replication programme (Drouin et al, 1990) and others a more gradual switch from early to mid/late S phase (Farkash-Amar et al, 2008, Strehl et al, 1997). Part of this discrepancy appears to be related to species and cell-type specific differences. Recent experiments by Katsuno et al (2009) have shown that origins within early and mid/late replicating chromatin have distinct affinities for different cyclin/CDK complexes, which may account for their differential activity during the replication programme. Their experiments in mouse embryonic fibroblasts identified the increasing activity of the Cyclin A2-Cdk1 complex from mid S phase onwards, with the complex appearing to be a regulator of origin firing in late S phase. The depletion of Chk1 (a known regulator of origin density) resulted in increased cdc25A expression and the hyperacetylation of Cyclin A2-Cdk1, which lead to the abnormal activation of late origins in early S phase. Given the importance of knowing the level of temporal control origin firing is subjected to, a more detailed description of the early to mid/late S phase transition would be of value to modelling of the S phase progression.

vi) During the analysis of the results, it became apparent that several of the outputs for the model, such as the transitions times and the completion times for chromosomes, could be altered if factories can progress between chromosomes. Another useful experiment concerning the R/G transition would therefore be to study the transition timing in different chromosomes as separate entities. Equally, the completion times of each chromosome would also be of great use. Whether chromosomes can be treated as individual entities or whether they are simply segments of an interrelated system is a question of fundamental importance to the model.

vii) The model makes a number of assumptions concerning the structure of replicon clusters which may lead to slight variations in completion times. The current model assumes a random distribution of replicons using sampling taken across an entire genome. However, it would be more efficient to cluster replicons of similar length. It may also be possible that different
areas of the genome are prone to clusters of different length due to different DNA sequence and origin densities (Cadoret et al, 2008). There is also the possibility that longer replicons lead to increased replication speed (Conti et al, 2007) so as to prevent replication delay. However, it should also be noted that factories have the potential for gradual assembly and disassembly (Sporbert et al, 2002), a process that will allow units of a factory that have completed replication to enter of soluble pool of components that is then able to support factory assembly at new sites. The gradual disassembly of factories does however lead to the question of whether a partially constructed factory can begin replication at some sites.

4.4 Perspective

Using these models as a first step towards the modelling of the mammalian cell S phase, a range of issues have been highlighted which can be considered further in subsequent studies. Given the limitations summarised in section 4.3, two of the main problematic factors are the concept used in the model to represent replication factories and the use of the R/G transition as a judging factor. Future iterations of the model should therefore explore these concerns, perhaps with a reconsideration of the model structure.
5.0 - Refinement of the Model

This chapter of the study seeks to build on the conclusions of Chapter 4 in order to develop a more detailed model which can be used to identify specific parameters which are particularly influential on the simulated system. Many assumptions were made with the previous models which need to be verified, and where possible these assumptions should be circumvented through alternative modelling theory.

One of the greatest limitations of the previous model was the concept of viewing replication factories as the active component of the model. Due to the lack of knowledge concerning these structures, attempting to parameterize their behaviour proved difficult. Without this information to fine tune their behaviour, factory dynamics become too simplistic and lead to an overly predictable and reliable system.

In this chapter, a second model is created which looks to an alternative method of cluster activation, inspired by lower level modelling of origins of replication. The limitation of activity by the number of factories is removed and is instead replaced by a desired level of replication activity which the system is optimised to produce, but the activity is not itself forced upon the model.

Additionally, the previous chapter highlighted the difficulty in evaluating the model through simulation of the R/G transition due to the lack of information concerning the exact nature of the progression. Whilst the transition could be approximated, the varied interpretations of this event within the field (which range from the 3C pause through to a graded transition) could allow a wide range of mechanisms to be optimal depending on the criteria chosen. This next generation model is developed with the aim of simulating an extended set of experimental observations which are more easily compared to outputs of the model.

The model developed in this chapter, along with the data derived from simulations, has been published under Shaw et al (2010). Experimental data for cluster activation dynamics provided by Dr. Apolinar Maya-Mendoza, as discussed in Maya-Mendoza et al (2010). The
TimEX-seq data used in both the estimation of total DNA replication and the final evaluation of the model was processed and analysed with help from Pedro Olivares-Chauvet.

5.1 Modelling Method

Firing Probabilities and Replicon Clusters

Previous models have often quantified activation of DNA replication through probabilities that express the likelihood that a pre-defined unit length of DNA will form an origin of replication over a given time period. This parameter is commonly expressed as I(t) and can be manipulated to represent a response to factors such as the increased availability of replication components or the local condition of the DNA. Used in many models concerning lower organisms (Herrick et al (2002), Rhind (2006), Goldar et al (2008)), the concept provides a quantitative measure for a process that is not entirely understood, integrating the many stochastic elements that make the system difficult to study.

The simplicity and flexibility of this concept may be ideal for the development of a model aimed at incorporating both the higher level structures of DNA and the wide range of factors that make the process more complicated in mammalian cells than in lower organisms. The most obvious and relevant structure to incorporate is that of replicon clusters. A replicon consists of the length of DNA replicated by a single origin. However, localized areas of DNA tend to fire their origins in synchrony, likely due to the local structure of the DNA. These replicons are likely replicated together within a replication factory, a nuclear body which contains the relevant replication machinery. For our modelling aim, this information implies that a cluster of replicons could be described as a single entity with a shared probability of firing (henceforth termed as the firing probability of a cluster). This probability can be manipulated with regard to relevant factors that will be introduced during development of the model, and provides an alternative mechanism to the use of replication factories as a manipulated component of the model.

One of the key goals of the model will be to determine profiles across S phase for these firing probabilities which incorporate the influence of relevant factors.
Limitation of overall activity

DNA replication in mammalian cells is not only a carefully orchestrated process, but it also has constraints governing its maximum activity. A number of different mechanisms could be in place to achieve this, the choice of which is likely to have drastic effects on the functioning of the model. Several possibilities are as follows:

i) The absolute number of factories is limiting; clusters would fire whenever there is a free ‘slot’ within a pool of factories that can assemble. This mechanism was tested extensively in the model studied in Chapter 4.

ii) Rate of DNA replication is limited; the amount of DNA replicated over a given time period is monitored and kept to a constant rate.

iii) The number of active forks is limiting; clusters would fire whenever there are sufficient numbers of components to build replication forks for all the origins within a cluster.

Each of these model concepts has its downfall. Use of an absolute number of factories or forks leads to low rates of overall replication during times of low fork rate progression in the variable fork distribution. As additional origins cannot be fired beyond a certain limit, the system has only limited ability to compensate. All three factors can also lead to overloading of cluster activation between different time frames; if many clusters are activated in time frame X in order to achieve a certain levels of activity, some of these will still be active in time frame X +1, hence leading to a reduction in cluster firing. This will depend strongly on the fork speed at the time of cluster firing and the distribution of firing within a time frame. It should also be noted that if fork rates are drawn directly from the distribution at each time point, there would be a decrease in overall activity during slow fork speed periods. This is also inconsistent with available data, as summarised in Fig. 5.2.7 which shows the sum of DNA replication over time drawn from chromosome 6 of the TimEX-seq data set from human ES cells (Desprat et al., 2009). Fork speeds will therefore be drawn from the distribution at time of firing and then remain constant.
As the output of replicated DNA is seen to be fairly constant, and its use as a limiting concept avoids the potential for an overall activity reduction, this will be selected as the factor that the model will seek to simulate.

The model will therefore be created with the following initial criteria and parameters:

i) The rate of DNA replication will be aimed at a constant level calculated by considering the total amount of DNA and the time that is available.

ii) Forks will progress at biologically relevant rates as estimated by the Takebayashi et al (2005) variable fork rate distribution.

iii) DNA replication will be aimed to complete within an 8-10hr time frame.

iv) The firing of origins of replication will be compared to existing profiles.

v) Parameters required to simulate the next-in-line mechanisms of S phase progression were taken from a recent study by Maya-Mendoza et al (2010) as summarised in Fig. 5.1. These parameters are representative of dynamics observed in nuclei during early/mid S phase.
Fig. 5.1 – The cartoon shows how the spatial architecture of replicon clusters within DNA foci contributes to S phase progression through a next-in-line mechanism. At the beginning of S phase (top) ab initio factory activation occurs at a sub-set of potential origins because of properties related to the chromatin context. After about 1 hour, forks between neighbouring replicons in clusters will meet and terminate by fork fusion. Replication activity then falls, and is compensated by new initiations. At this time, synthesis can spread by 3 distinct mechanisms: i) synthesis is activated in 1 neighbouring cluster and leaves 1 extending fork; ii) synthesis is activated in clusters on both sides of the primary active cluster or iii) elongation continues from the growing forks at the extremities of the primary active cluster. These 3 classes are formed in the following approximate proportions: 5:1:5 (Maya-Mendoza et al, 2010). How these growing forks relate to replication factories as replication proceeds is unknown. Initially, elongation continues at the expected speed. However, we see some evidence that forks eventually begin to stall, perhaps before their synthetic machinery is assimilated into a newly assembled replication factory at the neighbouring foci. Once S phase has begun, ~90% of initiation events are coupled to synthesis within previously active DNA foci.
Given criteria iv), the aforementioned features were built into a model that is based on human chromosome 6, to allow comparison with an existing study (Goldar et al., 2009).

**Model Implementation:**

Each iteration of the model involved the simulation of a linearised chromosome 6, using the method described in Section 4.3. This process involved the use of the following inputs:

i) The distributions of R- and G-bands for human chromosome 6 were taken from the March 2006 human genome assembly using the UCSC Table Browser (Karolchik et al., 2004).

ii) The architecture of replicon clusters was taken from data in Jackson and Pombo (1998).

Once activated, replication occurred using the variable replication fork elongation rates as detailed by Takebayashi et al. (2005). The estimation of the firing probabilities of clusters required to simulate an S phase that fulfils the aforementioned criteria is the subject of the first section of results. This was achieved by fitting to the natural S phase interval of 8-10 hrs. The firing probabilities are then applied to the linear framework, allowing the model to be tested against the specified criteria.
5.2. Results

5.2.1 Completing the Basic Framework - Calculation of firing probabilities

As a first step towards building the model, a set of firing probabilities was created which can be used to complete the parameterisation of the model framework. Given the imposed limit on the overall value of DNA replication over specific time period, a set of variable fork rates and an overall time, it is possible to estimate how many clusters are required per hour of S phase and the firing probabilities required to meet this.

i) How many replicon clusters are required?

The following data were gathered using 5000 simulated iterations of S phase over chromosome 6. Given the length of the chromosome in kbp, it is possible to calculate the fraction of the chromosome that must be replicated within each hour window of an 8 hour S phase. However, it must be considered that not all clusters of replicons will be activated at the start of each hour, hence a random start time within each hour was introduced prior to the calculation of DNA replicated per replicon. Whilst unlikely to be entirely accurate, it is not possible to estimate the speed at which clusters will begin to fire prior to the estimation of firing probabilities. The fork rate for each minute of these windows was drawn from the variable fork rate distribution which was smoothed to prevent large sudden jumps in fork rates. Once drawn, the fork rates remained constant for applicable replicons.

With these inputs, the number of clusters of replicons required to fulfil each hours DNA quota was calculated, drawing the number of replicons per cluster from the distribution described in Jackson & Pombo (1998). As mentioned previously, it must also be considered that, especially at low fork speeds, a replicon may not complete replication within the hour it began, hence some clusters contribute to the DNA replicated within a time frame without being counted towards activation events required within it. The number of clusters required is shown in Fig. 5.2.1.
ii) **Firing Probability Profiles**

To create the firing probabilities for replicon clusters, the absolute number of activations required per hour can be used to estimate the probability required per cluster per hour in order to achieve these values. The calculated firing probabilities have been extended to cover later time points by reusing the last estimated firing probability in the event that not all clusters have fired by the end of hour 8. The resulting data has also been divided into firing probabilities per cluster per minute, thus preventing sudden jumps between hour intervals. The calculated profile is shown in Fig. 5.2.

![Fig. 5.2](image)

Fig. 5.2 – White bars show the number of clusters of replicons required to fire each hour fulfil the quota of DNA replication for that hour. Red line shows the resulting firing probability, per cluster per minute, required to meet the expected cluster activation.

iii) **Verification of Data**

Chromosome 6 has been chosen for the current simulations in order to allow verification of the firing distributions. Goldar *et al* (2009) have created an origin firing profile describing number of initiations of DNA replication per time unit per unit length of unreplicated DNA. Given the current cluster firing profile, the model can measure the distribution of origin firing over chromosome 6 and compare it to the distribution estimated by Goldar *et al*. The aim was to recreate a tenfold increase in origin firing per given time unit per unreplicated cluster between 0 and 4 hours with a roughly linear increase. A steep decrease should then occur over the next 1.5 hours, with activity returning close to the original level of firing.
To perform this analysis, a linear model of chromosome 6 was created, as described in Section 5.1. The cluster firing profile, now describing the probability per cluster per minute of a firing event, can then be applied over a 12 hour time frame (allowing excess time for DNA replication to complete). 5,000 iterations of this process are run, with the number of origins fired and the amount of DNA replicated per minute being recorded and averaged (results shown in Figs. 5.3a and 5.3b).

Fig. 5.3a (left) – The blue line shows the progress of synthesis (DNA replicated in kbp/min averaged over 5,000 simulations) and black line the quota of DNA synthesis required to complete S phase within 10 hours. Red lines indicate one standard deviation either side of the mean.

Fig. 5.3b (bottom) – Average number of origins fired within 20 minute windows per unreplicated cluster. Number of unreplicated clusters is recorded at the first time point of each window.
While the results shown in Figs. 5.3a do give a close approximation of the DNA replication quota (having been optimised towards this end), the distribution of origin firing displays little resemblance to the results of Goldar et al (2009). The peak of origin firing shown in Fig. 5.3b is both too low and too late within S phase, with the total time required being consistently over 12 hours. If the two profiles are compared when scaled by maximum firing values, the simulated data is an average of +/- 45.7% from the experimentally determined profile. However, this result is to be expected as the only biological relevant imposition we have made is the organization of replicon clusters. Further factors therefore require consideration within the model.

5.2.2 Implications of chromatin

A limitation of the current cluster firing profile is that it considers all clusters to have the same properties and hence the same firing probability. This is of course untrue, as clusters will vary through many factors such as architecture, chromatin structure, gene content and GC content. The model created in Chapter 4 imposed the most obviously relevant factor to modelling the firing profile of clusters, the R- and G-banding (Schempp et al., 1978, Kim et al., 1975, Drouin et al., 1990). The model will also attempt to guide the activation of clusters through R- and G-banding via the incorporation of this information into the calculation of firing probabilities.

Whilst one can accept the general conclusion that R-bands should fire first, it is difficult to impose this knowledge onto the firing profile without knowing the aspect or aspects of the chromatin that are involved and how they affect each other. It is however known that late firing regions are likely to require the use of additional firing factors such as the Cyclin A-Cdk 1 complex (Katsuno et al., 2009). If this is therefore treated as an indication of G-band activity, a differential profile for R- and G-band clusters can be estimated. However, the extent of this differentiation still has to be defined.
From a modelling perspective, this question can interpreted in a number of ways. Firstly, the most extreme method of differentiation would be an absolute block on G-band cluster firing until all R-bands are completed. This would give the effect of the theorised 3C pause (Goldman et al, 1984), but relies on the cell accurately sensing the replication state of all R-Band clusters. A firing profile to reflect this would essentially be the current profile split down the middle in proportion to the ratio of R- and G-band DNA content. However, due to the effectively smaller pool of clusters to fire from, all firing probabilities would need to be proportionately increased. The enforcement of a block could also lead to an unnaturally long S phase unless the firing probabilities for R-Band clusters are maximized to 1.

Secondly, the preference towards R-Band clusters could be given by simply making their firing probability distinctly higher than that of G-Band clusters. However, this would require increasing the firing probability of G-Band clusters at a defined point either in time or in the replication program (i.e. when x clusters are completed) in order for them to activate within a biologically appropriate time frame. This style of approach leads to several questions however; should G-band clusters have a possibility of firing from the beginning of S phase, however small? How quickly should their firing probability increase? In what manner should this increase occur? How does the probability behave around a maximum point?

Given the model is to approximate G-band cluster activity via Cyclin A-Cdk1 availability, one solution is to approximate the introduction and maximization of the complex by scaling the *Xenopus* Cyclin A-Cdk1 observations of Katsuno et al (2009) into an 8 hour S phase. This would result in an introduction of the complex at 2 hours and a maximum probability reached by 6 hours with no indication of a decrease until after S phase is complete. In addition to this influence, the probability calculations must also take into account the constantly reducing size of the G-band cluster pool that is to be drawn from.

Two potential methods could be used to bridge the gap between the minimum and maximum points- either a linear increase in the Cyclin A-Cdk1 availability or an increase along a sigmoidal curve. Both were tested in order to find the optimum maximum probability as defined by the amount of deviation from the desired quota of DNA replication per minute. However, in order to determine this, the model first required an estimate of the firing
probabilities for the R-band clusters so that the distribution of DNA replication by R-band clusters can be calculated.

i) **R-Band cluster firing probabilities**

Given an 8 hour S phase, it is possible to estimate that at a constant speed of DNA replication, the time enquired to replicate all R-band clusters should be proportional to the amount of R-band DNA that occupies chromosome 6. The process described in Section 5.2.1 was used to calculate the amount of DNA replicated in different time frames by distributing the new quota (the R-band content of the chromosome) through the 3.28 hours allotted for the process. New firing probabilities were then calculated using the reduced pool of clusters (as only R-band clusters are considered).

After the calculation of adjusted firing probabilities for R-band clusters, it was possible to calculate the G-band cluster firing probabilities. Both linear and sigmoidal increases of the Cyclin A-Cdk1 complex were explored over a range of maximal firing probabilities for G-band clusters that were reached at hour 6. Once this target probability occurred, it was adjusted at subsequent time points with regard to the reduced pool of clusters. The overall number of clusters firing should therefore remain fairly constant. Without this compensation for the greatly reduced pool size, simulations would have the potential to extend indefinitely.

ii) **Linear Cyclin A-Cdk1 increase**

Using a similar method to section 5.1 iii), one can study the effects of the varied maximum firing probabilities. The firing probabilities for R-bands were drawn from the distribution created in Section 5.2.2 i), and the G-band firing probabilities were calculated from a linear increase from 0 at 2 hours to the variable maximum firing probability at 6 hours. The amount of DNA replicated over time at a variety of maximum settings is shown in Fig. 5.4.
The minimum variation from the DNA quota per minute was found to occur with a maximum G-band cluster firing probability of 0.324 per cluster per hour at the 6 hour marker.

Fig. 5.4 - Kilobase pairs of DNA replicated per minute using a range of G-band replicon cluster firing probabilities, with the maximum probability in each case plotted via a linear increase. Kbp of DNA replicated per minute is averaged over 1,000 simulations for each parameter. Colour of plotted lines indicates the probability of a G-band replicon cluster firing at the 6 hour time point with values as shown in the colour-bar key. Firing probabilities are measured per cluster per minute. Black line shows the quota of DNA to be replicated per minute. Green stars indicate the probability setting giving minimal variation from the DNA replication quota.
iii) **Sigmoidal Cyclin A-Cdk1 increase**

In order to perform a similar test for the sigmoidal increase, the shape of the curve also needed to be defined. A width of 26.2 was selected (giving 1% activity at 2 hours and 99% activity at 6 hours, centred around 4 hours) allowing calculation of G-band cluster firing via the following equation (A1 being 0 and A2 being the variable maximum firing probability).

\[
G\text{-band cluster Firing Probability}(t) = A2 + \frac{(A1 - A2)}{1 + e^{\left(\frac{(t-240)}{\text{width}}\right)}}
\]

Once corrected for the shrinking size of the pool of clusters, the probability profile was used to perform a similar analysis to 5.2.2ii) with results shown in Fig. 5.5.

Fig. 5.5 - Kilobase pairs of DNA replicated per minute using a range of G-band replicon cluster firing probabilities, with the maximum probability in each case plotted via a sigmoidal increase. Kbp of DNA replicated per minute is averaged over 1,000 simulations for each parameter. Colour of plotted lines indicates the probability of a G-band replicon cluster firing at the 6 hour time point with values as shown in the colour-bar key. Firing probabilities are measured per cluster per minute. Black line shows the quota of DNA to be replicated per minute. Cyan stars indicate the probability setting giving minimal variation from the DNA replication quota.
The minimum variation from the DNA quota occurred at the slightly lower 6 hour firing probability of 0.267 per cluster per hour. However, despite the likelihood of being the more biologically relevant method of Cyclin A-Cdk1 increase, the overall variation from the DNA quota was higher than in the linear model.

It should be noted in both these charts of DNA replication that the overall value often remains below that of the quota for the time unit, even at optimal settings. This is likely due to the earlier estimations of cluster firings throughout the hours having randomly distributed start times within the hour- however, as changing the firing probabilities effectively changes the distribution of firing within the hour slots and hence the number of clusters required, a more precise value would require a complex analysis. For exploratory modelling purposes however, these values are likely to be sufficient.

The firing probability profiles generated through the optimums of each technique (with adjustments after 6 hours to compensate for the dwindling pool size) are shown in Fig. 5.6.

![Fig. 5.6](image)

**Fig. 5.6 -** Firing probabilities generated after differentiation between R- and G-band replicon clusters and optimisation of G-band cluster firing probabilities. Red line- R-band replicon cluster firing probabilities. Blue line with squares- G-band replicon cluster firing probabilities calculated using a linear increase to an optimal probability of 0.324/60 per cluster per minute at 6 hours. Blue line with circles- G-band replicon cluster firing probabilities calculated using a sigmoidal increase to an optimal probability of 0.267/60 per cluster per minute at 6 hours.
5.2.3 Spatial architectures of replication foci

To this point the model had simulated the effects of replicon clustering within DNA foci, variable fork rates throughout S phase and the differential activation potential origins during early and mid/late S phase based on their chromatin environment. To add molecular complexity to the simulations, the model was now tested using different mechanisms of S phase progression (see Fig. 5.1). This aspect of the modelling is designed to assess how next-in-line and stochastic models of cluster activation influence S phase progression. In relation to the modelling of factories in Chapter 4, single-sided and bi-directional activations would mostly occur through neighbour activations (and more rarely by side by side stochastic activations) whilst \textit{ab initio} activations could occur as a result of the initial factory seeding or through stochastic jumps over at least one intervening cluster.

Simulations were performed using the conditions developed in Section 5.2.2 to test which activation parameters give the best fit to the established S phase duration (see Fig. 5.7). In this analysis, different modelling environments were compared using an end-time where 95% of DNA was replicated; this limits the effect of rare events that can lead to very long end-times. To simulate the effect of a next-in-line mechanism of origin activation different factors of increased activation (between x1 and x5000) were incorporated into the model. This feature alters the probability with which replicon clusters are selected for activation based on changes in the chromatin environment that arise during replication of neighbouring clusters. A low resolution scan of the parameter space, comparing a range of maximum firing probabilities for the sigmoidal curve (between 0.0001 and 0.0083 /cluster/min) highlights a number of regions of biological interest (see Fig. 5.7). In this phase plot, each of these areas of interest indicates the impact of different parameter sets and thus different mechanisms that are driving the system. In the examples discussed below, data was collected from 1000 simulations for each parameter set.
Fig. 5.7 - Using the method demonstrated in section 5.2, different firing probabilities were tested against a range of values to model spatial activation of DNA foci. As synthesis within active clusters completes, the extending forks growing out from the flanking replicons begin to interact with chromatin of neighbouring clusters. The figure shows the result of this influence when it can vary the probability of activation of the adjacent clusters - the extent of this increase was modelled over a range of probabilities from x1 (no change) to x5000 (highly probable). Given these parameter sets, an approximated phase space is created, which displays a number of key results: Black contours indicate completion times for replicating 95% of DNA. The red area indicates parameters giving a 95% completion time over 10 hours and the green area indicates parameter settings giving a 95% completion time of less than 8 hours. The white area therefore represents a set of biologically relevant parameters within which S phase would complete on schedule. To assist interpretation, additional features of interest have been imposed over the analysis: i) Magenta contours indicate the ratio of single:dual activation events and ii) Blue contours indicate percentage of ab initio firing events. Biologically interesting case studies A) – D) are indicated by coloured icons on the figure and discussed in the text.
The following conclusions were drawn from simulations that test 3 alternative models of S phase progression:

**i) Origin selection is stochastic (Positions A & B)**

A null hypothesis that ignores any relationship between DNA foci would simply alter the probabilities of cluster activation towards late S phase, based on expression of activating cyclin/Cdk complexes (creating results as seen in Fig. 5.5). In this case, the maximum probability of G-band firing defines the behaviour of the model. A maximum probability of 0.004/cluster/min was therefore tested as a case study (see Fig. 5.7 - position A). This parameter set gives an average variation from quota of 74.14 kbp/min and completes 95% of DNA replication within 8.4 hours, with absolute completion by 10.8 hours. The standard deviation at absolute completion was 67.0 minutes. This mechanism therefore provides a stable and timely completion of S Phase. However, this model does generate a high level of *ab initio* cluster activation of 41.5%, consistent with the stochastic firing mechanisms of the model in Chapter 4. Additionally, whilst the ratio of single sided firing events to dual sided is 2.2:1, this value is a consequence of the high levels of *ab initio* firing. Importantly, the distribution of origin firing in this case is skewed very late into S phase, and predicts a level of very late synthesis that is not seen experimentally. Predictably, increasing the maximum probability of activation results in a shift in activation but also leads synthesis to complete at unrealistically early times. However, the overall similarity between origin firing and the experimental profile of origin activation (Goldar *et al.*, 2009) is slightly increased by the imposition of the sigmoidal origin firing profile, with average differences of +/- 37.9 % and +/- 35.3% from the scaled experimental profile in each case respectively.

Allowing a small effect of fork elongation on cluster firing probabilities (with a maximum probability of 0.0033 /cluster/min and a x2 increase in cluster firing if forks are encroaching - Fig. 5.7 position B) reduces the length of S phase even though the maximum probability of activation is reduced. The variation of completion times is seen to rise slightly however, showing that slight spatial effects have little beneficial effects as far as the biological behaviour is concerned.
ii) **Encroaching forks drive cluster firing (Position C)**

Next-in-line models of S phase progression predict that the spread of encroaching forks is the driving factor that increases local firing probabilities. With regard to the model of Chapter 4, this mechanism would represent a nearest neighbour activation but with the potential for a bi-directional activation. To simulate this, the model was set with a low maximum value of 0.0008 /cluster/min for cluster firing probability and a high multiplier value of 1500, so that clusters with encroaching forks have a high probability of engaging synthesis (Fig. 5.7 position C). With these settings, 95% of DNA replication is completed in 8.3 hours and total completion occurs within 11.2 hours on average. The distribution of these completion times is more varied than in model (i), with a standard deviation of 73.7 minutes for the absolute completion times and 21.6 minutes for the 95% completion times. The ratio of single to dual activation is lower at 2.1:1, but many more clusters are operating through these mechanisms due to the lower *ab initio* levels. However, while the next-in-line conditions produce more variable end-times, the dynamics of cluster firing remain similar to the previous modes of activation, with timepoints varying by an average of +/- 37.5% from the scaled experimental profile, and yield a reduced rate (17%) of *ab initio* activation events.

It should be noted that the ratio of single activation events to dual can be increased towards the aim of 3:1 through reducing the maximum firing probability. However, this shifts the origin firing distribution later in S phase and increases overall variation in completion times. For example, a 3:1 ratio can be achieved through a maximum firing probability of 1500 and a firing probability multiplier of 0.01, but the peak of origin firing now occurs at 10 hours.
iii) Hybrid-driven cluster firing (Position D)

A final possibility is that alterations in cluster firing are driven by a mixture of the mechanisms explored in i) and ii). This was simulated in the model through a multiplier value for fork encroachment of 10 and a maximum firing probability of 0.0022 /cluster/min (Fig. 5.7 position D). The combination of factors still gives a 95% completion time of 8.4 hours with absolute completion in 11.0. The variation of the completion times lies between that of the two alternative models, as does the rate of ab initio activations at 31.0%. With a ratio of single activation events to dual activation of 1.86:1, these conditions allow a significant increase in activation by encroaching forks relative to the stochastic model. However, the spatial effects are not strong enough to drive a high ratio of single:dual cluster activation events, as is seen at higher levels of spatial activation by fork encroachment. Once again, the origin firing profile remains at a similar level of congruency to the experimentally determined profile, with an average difference of +/- 37.8% from the scaled experimental profile.

A summary of the replication profiles created by these mechanisms is shown in Figs. 5.8a and 5.8b. Fig. 5.8c shows an independent profile of DNA replication, created by experimentally derived data from a TimEX-seq data set from human ES cells (Desprat et al, 2009), which is comparable to the profiles displayed in Fig. 5.8b.
Fig. 5.8 a) - DNA replicated in kbp/min (averaged over 1,000 simulations). The colour of the plotted lines indicates parameters used in each set of simulations. Amplification factors (xn) define the adjusted firing probability that was applied when a cluster is activated by encroaching replication forks. Maximum firing probability refers to the probability of firing of a G-band cluster at the 6 hour time point, based on the optimal concentration of activating cyclin/Cdk complexes at that time.

Fig. 5.8 b) - Cumulative DNA replicated in kbp (averaged over 1,000 simulations) defined by synthesis within chromosomal R- and G-bands. Coloured lines indicate DNA replicated for 1 case study, with the solid lines indicating the total DNA replicated. Broken lines display DNA replicated after differentiation between chromosomal R- and G-bands, with dotted lines indicating replicated R-band DNA and the dashed lines indicating the replicated G-band DNA.

Fig. 5.8 c) - The in silico simulations shown in b) were tested against experimentally derived profiles using the TimEX-seq data set from human ES cells (Desprat et al., 2009). The replication profile for chromosome 6 was generated by segmenting the published TimEX-seq data into 100 time intervals and generating a cumulative frequency of genome duplication across the sample. Plots showing S phase progression were generated for the entire chromosome (Total DNA) and with differentiation between R- and G-bands (the distribution of bands being mapped as previously).
To explore how changes in the chromatin environment might influence the switching of synthesis between neighbouring replicon clusters, simulations were performed that incorporated sub-optimal fork elongation rates in order to mimic possible fork stalling, which might occur as synthesis switches from one replication cluster to the next. Variable probability settings in the range 1-50% were used to simulate different extents of fork failure. From these simulations, it is evident that the ‘fork elongation’ model is most susceptible to fork failure. Even so, a 6% probability of failure is required to drive completion of 95% of DNA replication beyond 9 hours, and a 16% probability of failure is required to delay beyond 10 hours (therefore becoming unviable). The ‘hybrid model’ is less sensitive to fork failure and completes 95% of DNA replication within 9 hours even with a 15% chance of fork failure and tolerating up to a 34% chance of failure whilst completing within 10 hours. Interestingly, increased levels of fork stalling also drives the hybrid model to generate a higher ratio of single:dual coupled activation events, while the spatially driven model maintains a constant ratio.

Predictions of replication timing profile generated by the final model were tested for biological efficacy by comparison with timing profiles generated using TimEX-seq protocols from human ES cells (Desprat et al., 2009). The replication timing for chromosome 6 was segmented into 100 time windows and a cumulative frequency profile showing the progress of DNA synthesis Figure 5.8c. The whole chromosome profile was then segmented into R- and G-band regions using the recognised coordinates (those used to classify R- and G- band clusters) to generate separate timing profiles for the two major chromatin compartments. Comparison of the S phase simulation with the TimEX-seq profiles (Figs. 5.8b and 5.8c) shows that the replication timing data generated from human ES cells has the same general characteristics as data generated by our in silico simulations. Similarities were most evident at the level of total synthesis, where in both cases the accumulation of replicated DNA was essentially linear. However, the individual profiles for replication of R- and G-band DNA show significant discordance. Based on these profiles, the basic assumption that synthesis of R- and G-band DNA occurs during mutually distinct periods of S phase appears to be flawed. Hence, while
the preference to engage synthesis in R-bands before G-bands is clear, the data do not suggest that an obligatory mechanism ensures that the cytologically defined chromosomal bands are replicated in a strict temporal order.

5.3 Conclusions

The model developed and discussed in this chapter was designed to allow two main aims - the allow observation of cluster activation under a variety of mechanisms and to allow closer study of the guidance capacity of R and G-bands with regards to these mechanisms. These, and other noteworthy observations, will be discussed below.

i) Cluster Activation Behaviour

A primary goal of this study was to observe the behaviour of models of S phase that were based on both the stochastic activation of replication domains and the sequential activation of genetically linked DNA foci, according to the ‘next-in-line’ hypothesis of S phase progression (Manders et al, 1992; Sporbert et al, 2002). As an alternative to these extremes, this study considered a hybrid variant that incorporates a combination of S phase propagation using the next-in-line principle together with a level of external or ab initio activation events that are not influenced by encroaching forks from neighbouring replicons. Such initiation events might arise with different probabilities at different times of S phase, for example in response to changes in expression of specific cyclin-Cdk complexes as S phase proceeds (Katsuno et al, 2009).

At one extreme, the stochastic model of activation leads to unrealistic levels of ab initio activation of clusters. With 41% of cluster activating by this manner, one would be very unlikely to observe such associations as those shown by Maya-Mendoza et al (2010). The comparatively high ratio of single to dual sided activations is due to the overall lack of
nearest neighbour activations, leading to a skewed ratio. Evidently, without a mechanism to
guide cluster activation in a more localised fashion, the simulated dynamics are a poor match
for those observed in vivo.

The ‘next-in-line’ model of activation, driven by encroaching replication forks, provides more
realistic dynamics, with ab initio activation near to the experimentally predicted level. This
allows a meaningful interpretation of the single:dual cluster activation ratio now that the
number of these events combined occurring is also in line with in vivo observations. Further
increases in the severity of this method of activation can shift this ratio to the observed
results, making this method a very suitable mechanism for cluster activation. Whilst this
method does provide the poorest approximation of the R/G-band dynamics, this may not
necessarily detract from this suitability (see Section 5.3 ii). It should also be noted that the
activation dynamics of Maya-Mendoza et al (2010) were observed during early/mid S phase
and therefore may not be entirely representative of the summed dynamics of the whole S
phase.

Finally, the hybrid model incorporates a spatial component and temporal features related to
changes in the chromatin environment. This results in an S phase driven by a low probability
of clusters firing within G-bands, but is then enhanced by the presence of encroaching
replication forks. Interestingly, the hybrid model is less reliant on fork elongation than the
basic fork encroachment model, yet it shares some of the spatial dynamic benefits whilst
being less susceptible to fork stalling. The fitness of this model is thus at least partially
reliant on the probability that forks progress from one cluster to the next.

As one would expect, the shared traits of the two previous mechanisms result in intermediate
values for the observed dynamics. The combination of the two mechanisms therefore appears
to provide little benefit to the mechanism over the ‘next-in-line’ model, other than being
more reliable in the case of high fork stalling.
ii) R/G-band inconsistency

The results indicate an obvious discrepancy between experimental data (Desprat et al, 2009) and the simulations of replication timing for designated R- and G-bands. This failure of the model implies that the chromosome-wide timing and order of R- and G-band replication is not defined with high precision. In particular, it is notable that while early cytological studies described a clear temporal separation in R- and G-band replication (Drouin et al, 1994) genome-wide analysis of the timing program has shown that R-bands replicate before G-bands but that replication of the cytologically defined DNA compartments occurs throughout S phase (Desprat et al, 2009).

Based on this analysis, it is clear that the temporal restriction of R- and G-band replication to specific periods of S phase is an over-simplification that must be re-evaluated if we are to develop biologically robust models of S phase progression. Specifically, it will become necessary to move away from the low resolution cytological chromosomal banding patterns, which generally incorporate chromosomal sub-domains of 5-20 Mbp, and towards high-resolution patterns of chromatin epi-states that better reflect local patterns of gene expression. Such improvements in resolution should provide a better insight into the molecular mechanisms that drive the spread of synthesis through mammalian genomes to ensure that synthesis is completed with the desired efficacy.

iii) Effects of variable fork rate distribution

Application of the variable fork rates to the model created in Chapter 4 led to a slight variation in cluster completion, but had little impact on the behaviour of the model or the main observations drawn from it. However, the application of the variable fork rate distribution to the model developed in this chapter has a significant impact on the firing probabilities assigned to clusters due to the requirement of the model to aim for a constant level of replication.
The variable fork rate distribution used in this model has the limitation of only covering 7 hours of S phase, hence an extension of the distribution was required to complete each simulation. The minimum for rate of 0.52 kbp/minute is extremely low especially when one considers that fork rates in the hours surrounding this measurement are also below 1 kbp/min; a study of fork rates by Conti et al (2007) in primary keratinocytes measured only 24% of replication forks as being between 0.5 and 1 kbp/minute. Given that 42% of the hour intervals measured in the variable fork rate study fall into this category, either there is a potential discrepancy between fork rate data or the distribution of fork rates is varied in different cell lines.
5.4 Perspective

The model discussed in Chapter 4 predicted that an optimum method of cluster activation would likely be a hybrid of the stochastic and ‘next-in-line’ models. The results above would indicate that this is the case, but only with a very slight tendency towards stochastic activation, which prevents the excessive activation of isolated clusters. Such activations could lead to large numbers of extraneous forks progressing through the genome, creating numerous regions that would be prone to damage and recombination due to potential fork failures. Despite this possibility, enough stochastic activation must occur to prevent large gaps in areas of G-banded clusters.

Analysis of the resulting data has called into question the suitability of the guidance of the model through the application of R- and G-bands. Given the insufficient resolution of banding data, this may either be due to a number of possible reasons. Firstly there may be a number of smaller bands within the currently recognised bands that have not been identified due to the resolution of the data. Secondly, replication may have been incorrectly approximated to R- and G-bands, and is instead more comparable to another factor that shares a distribution common to the R/G-banding patterns. Finally, the analysis may be being complicated due to the classification of R- and G-band clusters as being one or the other, whereas biologically there is likely to be a gradient of chromatin states between the two. Whilst the general trend of R/G-banding to replication timing remains, the data is insufficient to draw conclusions from, and certainly cannot be used to judge the suitability of mechanisms of driving S phase progression at this time.

The variable fork rate distribution was seen to have a great impact on the model, yet the current distribution is not necessarily suitable for the modelling of the entire S Phase. A complete distribution would therefore be a useful input for the model. This could be determined experimentally though the staining of a cell population with BrdU to determine fork rate whilst recording the progress of the cells in question through their S phase.
6.0 Data Acquisition

Previous chapters have indicated the need for a number of new data inputs required for more accurate modelling. The results drawn from the model developed in Chapter 5 lead to the conclusion that the use of R/G-banding data was insufficient as a guidance factor and as a judge of S phase accuracy. A new measure of S phase progression is therefore needed. Additionally, the significant effects of the variable fork rate distribution have lead to the requirement of a fork rate profile that covers the 10 hours of the mammalian S phase. This will in turn lead to the possibility of studying an origin density profile as discussed in Sections 6.2 and 6.3.

6.1 S Phase progression through replication patterns

Previous analysis has indicated the unsuitability of R/G-banding for the analysis of temporal S phase data. A different temporal guideline spanning S phase is required, allowing for comparison of S phase progression in order to produce temporally distributed data, and for comparison of the data produced by modelling. This study will explore the uses of the S phase patterns of replication towards fulfilling these aims. As these patterns are highly reproducible, they provide an excellent marker for the progression of cells into and through S phase.

One limitation with study of these patterns is that they are currently observed as sections of a three dimensional nucleus, whilst modelling within this project has occurred only in a linear manner. An analysis of these patterns is essential as a marker of the S phase for the timecourse experiments in Section 6.4 however, and there is the potential for recreation of the patterns through the manipulation of the linear model into a more biologically appropriate structure (such as the coiling of the linear chromosomes into a model nucleus).

In order to use these patterns to judge the progression of cell populations through the S phase (such as is required for Methods 3.3 and 3.6), large numbers of nuclei will need to be
labelled, identified and their patterns recorded. To this end, and in order to better understand the patterns displayed through S phase, efforts were made to automate the identification of nuclei and their replication patterns. MRC5 cells were used in this study, as their stable chromosome number may give more reliable patterns than cell lines with more varied nuclear states.

Populations of MRC5 cells were labelled and stained using the protocols in Methods 3.1, 3.2, 3.3 and 3.5 and images of the populations (as shown in Fig. 6.1) were collected using a Zeiss confocal LSM 510 META system. Software was then written for the ImageJ image analysis software (Abramoff et al., 2004) with the aim of scanning each nuclei as an individual entity which could then be identified using a Matlab code that would be written for pattern recognition. The overall process was to be formatted into a pipeline, with raw images being entered and the resulting statistics being outputted in Matlab.

Fig. 6.1 – Example output of an experiment involving the dual staining of whole nuclei. Cells were first pulsed for 30 minutes with BrdU and then pulsed with biotin for 20 minutes. Red channel shows nuclei staining resulting from the BrdU pulse, visualised through immunostaining with Cy3. Green channel show nuclei staining resulting from the biotin transfection visualised through immunostaining with Alexa488.
6.1.1 Nucleus Identification and Scanning

i) Nuclei Detection and Location

The first step in the creation of the pipeline was to enable the isolation of each nucleus as a separate entity. Whilst ImageJ does contain applications for identifying image components (Particle Analysis), they proved unreliable in identifying entire nuclei from the raw images, often isolating only parts of the nuclei. Background staining often gave false positives, where particularly highly localized areas of staining could be large enough to be highlighted by the Particle Analysis. In contrast, the area occupied by the staining of nuclei in late S phase was often too low to be identified and was ignored.

A series of additional processing steps were performed prior to the Particle Analysis in order to mitigate these problems. In order to locate the nuclei accurately, a series of despeckle commands were used to remove background and lower the overall image intensity. This removed all but the most intensely stained areas, allowing some differentiation between late nuclei and background. The image was then blurred (spreading the remaining stained areas to increase their number of pixels occupied) and the intensity increased, effectively amplifying the areas of highest staining. The process was then repeated, resulting in a high intensity approximation of each nuclei’s shape and size, with the pixels of late nuclei merging together to give a more solid shape (see Fig. 6.2). Additional rounds of despeckling were also performed in cases of high overall intensity of staining (which often resulted in high background).

Fig. 6.2 – Red channel from Fig. 6.1 shown after processing. The stained nuclei form more solid, isolated images that are easier to identify through particle analysis.
Particle analysis could then be performed on the image, set to identify shapes between appropriate size ranges (100-300 pixels in figures shown, but adjusted for magnification) and fit an ellipse to the selected shape. The size specification allowed differentiation between nucleus sized objects and the few remaining areas of coalesced background.

It should be noted that with the background staining removed from the image, overlap of the nuclei proved not to be a common issue. Cells used in the experiments were also only grown to sub-confluent levels (70% confluency unless otherwise stated). In the rare cases of overlap caused by the blurring steps, both nuclei would be removed by the size thresholding for particle analysis. However, the loss of a few close nuclei in such cases was minimal in comparison to the gain in late nuclei detected through over-amplification, hence this process was retained.

The result of the particle analysis process was a list of the nuclei locations as a series of ellipses, described by their centre points, major and minor axes and their angle of tilt.

**ii) Nuclei Scanning**

A number of methods were then used in an attempt to scan the nuclei, with cross sections of intensity being taken at various angles across the nuclei. However, the resulting data often missed many of the subtleties that would be required in order to identify between the different replication patterns. So as to record as much information as possible, the area of the ellipse estimated by the particle analysis was scanned in its entirety. However, as the intensity scans could only be performed as a linear process over a 1 by X series of pixels, multiple scans were to be taken, recorded and then organized with respect to each other. The information describing the ellipse was used to generate a list of scanning co-ordinates, with an appropriate offset of null intensities added to either end of the pixels scanned so that the ellipse retained its shape within an artificially created grid (see Fig. 6.3).
Fig. 6.3 – Nuclei Scanning. An ellipse surrounding the nuclei is recreated and then converted into a list of horizontal pixel co-ordinates. These are then scanned as rows (shown in red, the yellow arrow indicating scan progression). Null intensity ‘pixels’ are added (blue) to retain ellipse shape.

This information was stored with line breaks where appropriate so that it could be recreated during the Matlab analysis along with its paired scan from the second image. The Matlab heatmap image of the result of the process is shown in Fig. 6.4.

Fig. 6.4 – Left- A nuclei displaying the mid S phase replication pattern (cell labelled with BrdU and stained with Alexa 488, visualized on the Zeiss Confocal microscope). Right- A heatmap image created in Matlab using the data from the intensity scans.
iii) Recreation of Nucleus Boundary in Matlab

Following the scanning of the data into Matlab as a grid of intensity values, the area of focus (i.e. within the nuclear boundary) had to once again be established. To facilitate interactions with other Matlab tools, a Matlab package for creating ellipses (fit_ellipse, available from the Matlab file exchange) was used to estimate the area the nucleus occupied. To ensure that the package did not calculate that the ellipse extended beyond the grid, a border of 20 null intensity pixels was added around each image. The area highlighted was then tested through a series of techniques aimed at accurately determining the staining pattern inside.

6.1.2 Automated Replication Pattern Recognition

After consideration of the resulting images, a number of difficulties in their comparison immediately became apparent:

1) Imaged nuclei were often different sizes (owing to both variation in cell size and the orientation of the nucleus as the cell was stained and imaged).

2) The intensity of staining varied greatly both within a single dataset (i.e. images taken from a single slide) and between datasets (i.e. images taken from different drug concentrations and different experiments).

3) The simplified classifications of early, mid and late S phase are not distinct subphases. As the cell moves from one state to the next, intermediate patterns become evident.

4) Use of different staining reagents (e.g. BrdU, EdU and Biotin) lead to slightly varied patterns (likely due to different degrees of incorporation and variation in antibody/chemical labelling affinities). The textures of the resulting patterns therefore also varied.
As a result of these problems, the ideal solution required tolerance for the different staining intensities and some degree of flexibility in the definitions of the different classifications. A cut-off point between the different subphases needed to be imposed, in an attempt to replicate the judgments made when judging the patterns by eye.

To this end, the development of this tool involved iterative testing with a number of datasets. The first was a manually annotated dataset, used to generate rules that attempted to differentiate between the staining patterns. These rules were then tested on a variety of datasets that demonstrated a range of staining types and the results were compared to manually annotated dataset.

i) **Image descriptors**

In the first iteration of testing, a wide range of measurements were taken describing different aspects of the image intensity and the relationship of pixels to their neighbours. The aim was to find a subset of these descriptors that could be used to give accurate differentiation between the images using cut-off points along these key measurements. The different modes of analysis tested were:

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Analysis entailed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intensity</strong></td>
<td>The sum intensity of the image divided by the number of pixels</td>
</tr>
<tr>
<td><strong>Pixel Islands</strong></td>
<td>The number and length of groups of pixels over a cut-off intensity</td>
</tr>
<tr>
<td><strong>Masked Data</strong></td>
<td>The number of pixels above a number of different cut-off intensities across the image</td>
</tr>
<tr>
<td><strong>Deviation</strong></td>
<td>The sum of the difference between each pixel’s intensity and the mean intensity</td>
</tr>
<tr>
<td><strong>Background Intensity</strong></td>
<td>Recorded by ImageJ with regards to the entire field of view (i.e. prior to identification of individual cells)</td>
</tr>
</tbody>
</table>

Further measurements were made using the Matlab image analysis functions. A Grey-Level Co-Occurrence Matrix (GLCM) was created to allow examination of the image texture.
matrix represents a measure of how often pixels of certain intensities are next to each other.

The following descriptive parameters are then generated from this data:

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Analysis entailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrast</td>
<td>The measure of difference between each pixel and its neighbour</td>
</tr>
<tr>
<td>Correlation</td>
<td>The measure of how correlated a pixel is to its neighbour</td>
</tr>
<tr>
<td>Energy</td>
<td>The measure of how constant the relationships between different pixel intensities are (i.e. is a pixel equally likely to be next to a pixel of any other intensity?)</td>
</tr>
<tr>
<td>Homogeneity</td>
<td>The measure of how the relationship between pixels of the same intensity compares to that of different intensities</td>
</tr>
</tbody>
</table>

Each of these values (including values scaled by the intensity of the background) were then plotted against each other in an attempt to differentiate between a set of 120 sorted patterns (30 early, 30 mid, 30 late and 30 unstained) selected from both BrdU-Cy3 staining and Biotin-Alexa488 stained populations. The sorted results of all combinations of 2 and then of 3 different variables were compared. A typical result is shown in Fig. 6.5.

Fig. 6.5- An example population of 120 cells sorted by staining intensity and average pixel island size.

Red circles- Early S phase nuclei.

Yellow stars- Mid S phase nuclei.

Blue crosses- Late S phase nuclei.

Green stars- Unstained nuclei.
However, even upon complete separation of a training set, each attempted combination of rules failed to correctly identify over 70% of staining patterns in the test set. The variation that existed in the stained cell populations proved too dominant for separation using any combination of the analytical rules.

**ii) Void measurements**

Given the inability to find correlations between replication patterns and measures of intensity and pixel relationships, a second method was tested, whereby the distribution of areas that lacked staining (void pixels) was tested. Initial analysis involved measuring the average distance and distribution of non-stained areas from the centre of the ellipse that defines the area of the nucleus. Fig. 6.6 demonstrates the steps involved in this process.
Nuclei were described by the total number of void pixels (i.e. the lack of staining) and the variation between the concentric rings. An ideal separation of the staining patterns would theoretically occur through the observation of high numbers of void pixels in late nuclei, medium numbers in mid nuclei and low numbers in early nuclei. Any unusually high staining in mid nuclei would be countered by the high levels of variation between the different rings.

Fig. 6.6 – The stages of processing used in measurement of the void pixels. Axes indicate number of pixels.

Top plot shows the raw data (a scanned mid S phase nuclei).

Middle plot shows the mask of the nuclei, with areas of void pixels in blue and areas with staining shown in red. Arbitrary thresholds for cut-off values were tested in order to retain as much descriptive information as possible (red border excluded from the analysis).

Bottom plot shows the imposition of rings over the nuclei. The number of rings was varied in different calibration tests so as to obtain the optimum resolution.
(due to the characteristic lack of staining between the nucleoli and the nuclear periphery in mid S phase nuclei).

Fig. 6.7 shows a plot of the results of this analysis. The general trends predicted are correct, with low levels of void pixels seen in early, and high levels in late. Variation between rings is also higher in mid and late nuclei than is observed in early nuclei. However, once again, the variation between nuclei of the same pattern category is often higher that that between the categories. Whilst the technique identifies the general trends, it is insufficient to sort mixed populations accurately.

It was noted that the use of concentric rings essentially removes a dimension from the data, as isolated formations of DNA replication (e.g. a large replication zone in a late nuclei) are averaged across the ring. This may remove some descriptive information that could be used to differentiate between the staining patterns, given the often ‘spotty’ texture of late cells.
iii) **Texture Analysis**

As a result of the observations made in section 6.1.2 ii), another analysis method was created to measure the texture of the staining. The concept was to differentiate between large areas of relatively solid staining (as observed in early nuclei and in segments of mid nuclei) from relatively spotty staining (as observed in late nuclei). To this end, grids were imposed over the specified nuclear area with two different window sizes and the pixel intensities in each grid square were averaged. A larger grid size would give a blurred image of the staining of nuclei, whereas the smaller grid would record fine detail. In their comparison, one could then determine between a solid image, which would have similar readings across each grid, and a spotty image, which would have occasional peaks in the high resolution grid that would then be smoothed in the low resolution grid.

The variation between staining again led to failure of accurate identification however. As shown in Fig. 6.8, staining of nuclei can occasionally lead to spotty images even in early nuclei, and even with adjustment for intensity the textures can be difficult to differentiate.

Fig 6.8 – Nuclei labelled with BrdU and immunostained with rat anti-BrdU and goat anti-rat Alexa 488. From left to right- early, mid and late replication patterns showing similar texture.
6.1.3 Observed difficulties of automated pattern recognition

The various methods of analysis of the image data has illustrated a number of issues that must be overcome in order to create reliable automated identification algorithms.

From the initial datasets, it is evident that the method of staining of nuclei leads to large variation in the resulting images. Fig. 6.9 shows three mid S phase nuclei, stained by three different methods. The differences between the images, introduced by the staining methods, are potentially more dominant than the differences between nuclei displaying different S phase patterns. The identification algorithm would likely need to be calibrated separately for each type of staining.

![Fig. 6.9 – Mid S phase nuclei stained via three different protocols. From left to right: Biotin transfection (Methods 3.2), BrdU staining (Methods 3.1), EdU staining (Methods 3.3).](image)

Significant variation also occurs within the different S phase populations of similarly stained cells. This often occurs as a result of the intensity of the staining, which can vary within in a single population. A high degree of staining can lead to the appearance of active foci being larger and blurring with those nearby. This has the potential for creating areas of a mid S phase nuclei that can be mistaken for an early S phase nuclei. General trends are seen in intensity, with cells local to each other having similar levels of staining. However, over a single cover slip the intensity can have localised peaks. This effect can be mitigated through measurement of the background intensity and the lowering of the image intensity overall, but this may result in the loss of relevant data, especially with regard to late S phase nuclei.
Methods of analysis that relied on the comparison of the nuclei to an expected pattern (such as the sequential ring analysis designed to specifically differentiate mid S phase cells) suffered from the range of sizes and shapes of nuclei that occur across a population. Other forms of image analysis overcome these problems through the use of landmark features. The nucleus is too fluid an environment to allow this process. The nuclear boundary can be used as a border to stretch the nucleus into a template shape, but the occurrence of indentations in the overall shape of the nucleus tends to make this process unreliable and leads to the overrepresentation of areas that are stretched more than others.

Each of the methods of analysis seeks to draw a boundary between elements of the test population given a set of criteria. However, the progression of the cell cycle itself is continuous, with the patterns progressing from one to the next. This can result in patterns that display either two partial patterns laid over each other or the incorporation of two separate patterns in different areas of the nucleus. Identification of the dominant pattern often relies on the comparison of multiple criteria that describe each of the S phase patterns—hence correct identification of a ‘hybrid’ pattern relies on the criteria being weighted to represent the dominance of each pattern to the correct extent. This quantification process requires extremely careful balancing to prevent all nuclei that could potentially be a certain pattern from being selected over a more dominant pattern.

Whilst each of the methods tested could differentiate trends between each S phase subpopulation, the boundaries between the populations were never clear. The net effect of all aforementioned sources of variation is to create a wide distribution within each descriptive measurement that can be derived from the image. This variation is often so great that it causes overlap between the subpopulations at the trend boundaries, which renders the techniques too unreliable to successfully identify the full range of patterns.
6.1.4 S Phase progression measurements

Despite the perceived difficulties in automated recognition of S phase replication patterns, the study has provided a large amount of data in the form of the manually annotated datasets. These counts describe the distribution of the MRC5 test population over their cell cycle, and combined with a measurement of the length of the cell cycle, this data can be used to estimate the duration of early, mid and late S phase for the test population.

The duration of the cell cycle was measured through a study of the doubling time of the MRC5 population. Cells were seeded on cover slips at low densities and the population counted every 4-10 hours until the cells became confluent. After a lag phase in cell growth, the doubling time for the populations was averaged to 22.5 hours once the cells were in the exponential growth phase.

Counts from the manually annotated datasets were then used to estimate the fraction of the cell cycle that cells occupy early, mid or late S phase. These values gave an estimated 7.0 hour early S phase, 2.14 hour mid S phase and a 1.55 hour late S phase. The sum of these values equates to an S phase of 10.7 hours, which is similar to previously observed measurements of S phase duration.

6.1.5 Conclusions

The development of automated identification has proved to be a complex process. A number of difficulties in the process have been identified; however, the solving of these issues is unlikely to be feasible within the timescale of this study.

The observed data has provided estimates of the duration of early, mid and late S phase however, which will be required for the analysis of data in Section 6.4.
6.2 Manipulation of replication forks and origin densities

The evident importance of replication fork rates has prompted the creation of a profile covering the entirety of S phase in order to allow more accurate modelling. Additionally, the study of S phase progression was conducted in MRC5 cells, whereas the previously used variable fork rate profile was measured in HeLa cells. As there is the potential for different behaviours of replication in different cell lines, a novel replication fork rate distribution is required.

However, prior to the study of fork rates over the S phase, a number of trial studies were performed. The focus of this was to test out the extremity of a separate hypothesis which could potentially have an equally large impact on the modelling of S phase. A number of publications (Marheineke & Hyrien 2004; Maya-Mendoza et al., 2007; Courbet et al., 2008) have demonstrated that the activity of replication forks and the density of origins within clusters are linked. In the case of replicational stress due to slowed forks (such as from aphidicolin treatment), additional origins can fire to keep overall replication activity at a constant level. One could hypothesise that during points in the S phase where replication forks move at slower rates, the average distance between origins may also be lower.

The mechanics behind this hypothesis can be aptly demonstrated through experiments involving aphidicolin and caffeine. Aphidicolin decouples the helicase and polymerase complexes in active replication forks, leading to slower fork progression. Caffeine acts from the opposite direction, inhibiting the ATM and ATR pathways (Maya-Mendoza et al., 2007) leading to the firing of additional origins. In each case, one would expect the responding variable (origin density for aphidicolin and fork rates for caffeine) to be manipulated in order to balance the system.

The measurement of fork rates and origin densities were performed using BrdU and biotin staining (respectively) of DNA during replication as explained in the Methods 3.1, 3.2, 3.3 and 3.4. Three replicas of each experiment were performed on MRC5 cells when populations had reached between 60 and 70% confluence.
6.2.1 Manipulation of Replication Fork Rates

To determine the effect of reduced fork rate on origin density, 60-70% confluent MRC5 cells were incubated for 12 hours with varying concentrations of aphidicolin. Cells were either treated with 2 µg/ml, 0.4 µg/ml, 0.2 µg/ml, 0.1 µg/ml, 0.05 µg/ml or no (Control) aphidicolin. After twelve hours incubation at 37°C, each cell population was pulse labelled with BrdU for 20 minutes or transfected with biotin for 30 minutes. The DNA was then extracted and the fibres immuno-stained. Figure 6.10 shows extracted fibres labelled with BrdU, demonstrating the efficacy of aphidicolin in reducing the rates of replication forks. No results were taken for the 2 µg/ml population as replication forks were effectively stalled, preventing any measurements of length from being taken.

Fig. 6.10 – Fibres extracted from three populations of MRC5 cells. Fibres were labelled with BrdU and immunostained using a primary rat anti-BrdU antibody and a secondary goat anti-rat Cy3-linked antibody.

Top – Fibres extracted from the control population (no aphidicolin).

Middle – Fibres extracted from cells treated with 0.1 µg/ml of aphidicolin for 12 hours.

Bottom – Fibres extracted from cells treated with 0.4 µg/ml of aphidicolin for 12 hours.
The reduction in fork speeds over the measured concentrations is shown in Fig. 6.11.

Further cell populations were labelled with biotin over the same range of aphidicolin concentrations and the distance between origins was measured. The results of this are shown in Fig. 6.12.

The reduction in fork rate is clearly accompanied by a reduction in the inter-origin distance. Importantly, this compensatory mechanism is seen to operate at fork rates that occur within the previous variable fork rate distribution. Previous modelling within this project has drawn replicon lengths from the same distribution independent of the fork rate in operation at the

Fig. 6.11 – The effect of aphidicolin on replication fork speeds. Bars indicate one standard deviation to either side of the data point. Data derived from BrdU labelled MRC5 cells, n = 38-1321 forks.

Fig. 6.12 – The effect of aphidicolin on origin density. Bars indicate one standard deviation to either side of the data point. Data derived from biotin labelled MRC5 cells, n = 27-46 origin pairs.
time of the origin’s activation. The results of the feedback mechanisms are demonstrated in Fig. 6.13. The increased origin density effectively cancels out the reduced fork rate, giving consistent replication times of around 30 minutes to replicate a single replicon. One outlier is the measurement at 0.4µg/ml, which shows completion times having risen to 50 minutes. This indicates a failure of the feedback system, where increased origin firing has peaked yet can no longer compensate for the continued drop in fork rates. The lowest value for fork rates which allows the system to continue to balance is ~0.3 kbp/min. Over 30 minutes, this would allow replication of 18 kbp of DNA, which is demonstrated in Fig. 6.12 to be close to the lowest inter-origin distance.

6.2.2 Manipulation of Origin Densities

To study the reversed mechanism of origin/fork rate feedback, 60-70% confluent MRC5 cells were incubated with varying concentrations of caffeine. Cells were either treated with 1mM, 2 mM, 3 mM, 5 mM, 7 mM or no (Control) caffeine. After a 1 hour incubation at 37°C, each cell population was pulse labelled as in 6.2.1. Figure 6.14 shows extracted fibres labelled with biotin, demonstrating the efficacy of caffeine in reducing origin density.
The reduction in origin densities over the measured concentrations is shown in Fig. 6.15.

Further cell populations were labelled with BrdU over the same range of caffeine concentrations. The results of this are shown in Fig. 6.16.
Once again, the reduction of origin densities correlates with a reduction in fork speed so that the overall system maintains a balanced level of total replication (see Fig. 6.17). However, unlike the previous experiment, the controlled variable can only be reduced to a basal level by the addition of caffeine, with no further reduction of inter-origin distance once a threshold has been reached. This plateau effect prevents that the system from reaching an unbalanced state as demonstrated in Section 6.2.1.

Fig. 6.16 – The effect of caffeine on replication fork speeds. Bars indicate one standard deviation to either side of the data point. Data derived from BrdU labelled MRC5 cells, n = 101-433 forks.

Fig. 6.17 – Time required for a pair of replication forks to replicate a single replicon at different levels of caffeine treatment.
6.2.3 Conclusions

A number of conclusions can be drawn from these experiments which are relevant to the development of further models. Firstly, the experiments demonstrate the ability of the system to balance perturbations in the controlled variables through manipulation of the responding variable in either direction. This is important as it indicates the potential that the reduction in fork rates seen at certain timepoints in the current variable fork rate distribution could be linked to a localised increase in origin density in the chromatin being replicated at those timepoints. As the system has been demonstrated to react in either direction, this possibility cannot be ruled out. A profile of replication fork rates over S phase should be accompanied by a profile of origin densities in order to test whether the changes in fork rates are due to origin density variations rather than other variables such as a response to the states of the chromatin itself. It could be postulated that origin densities are higher in areas of housekeeping genes in order to cause a reduction in fork rates so as to lower the chance of mis-replication.

It was also noted that the time required to replicate a single replicon was maintained by the system at around thirty minutes on average. However, the occupation and replication of a cluster by a factory is generally accepted as requiring 45-60 minutes in other mammalian cell lines such as HeLa (Jackson & Pombo, 1998) and C2C12 mouse myoblasts (Leonhardt et al, 2000). There are a number of possible explanations for this observation, dependent on the grouping of replicons. If clusters are formed of groups of variable length replicons, a cluster would need to contain a replicon of around 168 kb in length in order for the factory to be required for an hour. Whilst such replicons have been observed, they are unusual in the studies shown above, where the pooled control replicons have a mean length of 92.3 kbp, and a standard deviation of only 37.7 kbp. Only 4% of replicons are above the 168 kb threshold, which falls far short of the 1:3 ratio required (given an average cluster of 4 replicons). Alternatively, if clusters are formed of replicons of similar length, the additional time required could be due to the assembly and disassembly of replication factories, during which period no replication would take place. From a modelling viewpoint, this could have important consequences depending on the limiting factor that controls the overall activity of
the system. If total replication is limiting, little change would be required, other than the overlapping activation of slightly higher numbers of factories in order to keep the overall activity at the desired level. However, if replication machinery is limiting, the requirement of time to assemble and disassemble factories would mean that components have been removed from the pool of replication machinery without contributing to active replication for a substantial time period. The grouping of similar length replicons into clusters would imply that replicons within a cluster could no longer be drawn independently from a distribution of replicon lengths.

6.3 Synchronisation of Cell Populations

In order to create profiles of origin densities and fork rates from populations of MRC5 cells, the populations first needed to be synchronised so that the populations could be sampled at distinct timepoints within S phase. A number of methods of synchronisation were tested, with the aims of achieving close synchronisation whilst avoiding stressing of the cells or the alteration of variables relevant to S phase (with focus on the desired observational variables in particular). It should be noted that even with a perfectly synchronised population there is still likely to be variation in the distribution of cells throughout S phase when the population is released. This may be due to the progression of cells through S phase at different rates. Synchronisation of cells using hydroxyurea, performed by O'Keefe et al (1992) showed over an 80% synchronisation of cells within the first hour of S phase, yet the distribution of cells through their cell cycles still showed significant overlap as S phase progressed.

The results for three of the methods of synchronisation tested are shown in Figs. 6.18, 6.19 and 6.20. Cell populations were grown in T75 flasks until 70% confluent and then split into 24 well plates either before treatment (in the case of serum starvation) or after treatment (in the case of nocodazole treatment). At each hour of the timecourse, two wells were labelled with EdU, allowing observation of the distribution of cells throughout the cell cycle at that time point. A minimum of 350 cells observed from each cover slip, and these were summed to
give an estimated distribution of the overall population of that time point, as indicated by bars in each figure.

Fig. 6.18 shows the effects of serum starvation for 24 hours followed by release through the addition of fresh media (see Methods 3.7). Very low levels of synchronisation were observed.

A series of tests were then conducted using combinations of aphidicolin and nocodazole in an attempt to increase the synchrony of the population through the use of 2 sequential blocks in the cell cycle (see Methods 3.8 & 3.9). Cells were first incubated at 37°C with 100ng/ml nocodazole for 12 or 16 hours, which prevents the formation of mitotic spindles, hence preventing cells from progressing through mitosis. Blocked cells were harvested and then reseeded with 2 µg/ml of aphidicolin and incubated at 37°C for 12, 14 or 16 hours. This concentration has more significant effects than the lower concentrations used in Section 6.2, stalling the cells at the G1/S phase border. Cells were then released through washing and the synchrony of the population measured. The combination of incubations giving the best synchrony was 12 hours of nocodazole treatment followed by 12 hours of aphidicolin, which yielded close synchrony in addition to high numbers of actively replicating cells. The synchrony of cells through 4 hours towards the end of S phase is shown in Fig. 6.19.
As a preliminary study, DNA fibre analysis was then performed on the synchronised population to test for residual effects of aphidicolin. Despite multiple washes, the aphidicolin was shown to continue to reduce fork rates, making it inappropriate for the experiment. A 12 hour incubation of cells with nocodazole was shown to give a partial synchronisation of the cell population (see Fig. 6.20) without disrupting replication forks or slowing the cell cycle.

This method of synchronisation was used in the subsequent timecourse experiment. Whilst the population is only partially synchronised, there should be a majority peak which can be identified for each timepoint. The measurements were extended to 6-17 hours after nocodazole release so as to allow observations of the entirety of S phase.
6.4 Experimental determination of Fork Rates and Origin Densities for the MRC5 cell line

Using the protocol determined in section 6.3 for partial synchronisation of MRC5 cell populations without noticeably disturbing cell growth or altering measured variables, it was possible to create profiles of temporally variable S phase parameters. Two of these in particular will be very useful in providing accurate inputs for the next stage of modelling; these are a fork rate distribution and an origin density distribution.

6.4.1. Experimental Method

In order to create as closely synchronised profiles of fork rates and origin densities as possible, nocodazole was used to partially synchronise a large cell population which was then split into three population streams. Each stream consisted of 12 subpopulations, one for each hour of the measured timecourse.

The first stream was grown on cover slips in 24 well plates, and was used to measure the synchrony of the population and the progress through the cell cycle. This was measured using EdU labelling and whole nuclei visualisation (see Methods 3.3). Two subpopulations were labelled at each timepoint, allowing both an average to be taken (for increased accuracy) and to allow a comparison of synchrony of cells within the same timepoint. In previous experiments, this had always given very similar distributions of cells in the different sub-S-phases, indicating that populations tend to progress at the same rate despite being split into subpopulations.

The second two streams were each grown in twelve 25 mm plates. A single plate from each stream was labelled every hour. The first stream was labelled with BrdU (see Methods 3.1) and the second transfected with biotin (see Methods 3.2). The cells for these streams were harvested at the end of the timecourse and used to create DNA fibres for immunolabelling (see Methods 3.4). The data in Section 6.2 regarding the relationship between origins and fork
rates could then be used to analyse these two streams. If the variable fork rates occur in parallel to variable origin densities, one would expect the feedback system to compensate to a similar extent as summarised in Figs. 6.13 and 6.17.

6.4.2 Results

Prior to the measurement of DNA fibres to determine fork rates and origin densities, the distribution of the cells throughout their cell cycle was measured over the length of the timecourse. The results are shown in Fig. 6.21. The overall level of labelling was lower than was expected after the testing in section 6.3, indicating that some cells may have progressed slowly out of the block or may have died as a result of the protocol. Whilst the slowly progressing cells could lead to additional noise at later timepoints, dead cells could not influence proportions of cells in early, mid and late S phase as they would not be stained. The remaining cells appeared healthy and progressed at a normal rate. The partial synchronisation of the population is likely to have provided data representative of specific timepoints, with the S phase temporal fraction with the greatest dominance giving the strongest signal as it progressed. The analysed fibres for each hour interval will contain background noise from contamination of other temporal fractions, but this should still yield a peak value. It was also noted that very few cells were actively replicating in hour 6. This may imply that the majority of cells did not enter S phase until hour 7 or 8.

Fig. 6.21 – The distribution of cells throughout their cell cycle after partial synchronisation via 12 hours of nocodazole treatment. n = 700 cells per timepoint.
With the population partially synchronised, the fork rates and origin densities for each temporal fraction were then measured. Fig. 6.22 shows the distribution of fork rates for each point in the timecourse. The individual fork measurements have been binned into 0.1 kbp/min intervals and the frequency of each bin was expressed as a percentage of the total.

Fig. 6.22 – Profiles of fork rates for each hour of the time course. Plots indicate binned data, with each bin spanning 0.1 kb/min. ‘Hour’ indicates time after nocodazole release.
As expected, the data features a degree of noise due either to contamination of timepoints with other temporal fractions of variation between fork rates within fractions. If the median of each profile is taken, it is possible to produce a variable fork rate distribution that loosely follows that of Takebayashi et al (2005) (see Fig. 6.23). However, the plot does not appear to explain much of the data, with many median values lying around the mean of the profile, likely due to the more extreme values being masked by noise. A different method of analysis was required in order to analyse the data, allowing a more complete explanation of the individual temporal profiles. Such a method could then be applied to the distribution of origin densities, given that their profiles were generated from a similarly distributed population.

![Figure 6.23](image)

**Fig. 6.23** – The median values of the temporal profiles is used to create a variable fork rate distribution for MRC5 cells. The profile lacks the potential to explain the more extreme values observed in the temporal profiles however.

### 6.4.3 Analysis

In order to determine a method of analysing the replication fork profiles, the underlying structure of the population had to be determined so as to understand the level of interference that contamination from other timepoints could potentially cause. If the population was concentrated around a single peak, a true replication fork profile could be extracted from the data. However, if there were also a number of other peaks of similar intensity, or the distribution of the population was bimodal, analysis of the data would likely to involve too much variation to give a meaningful conclusion.
To study the underlying structure of the population, two methods of analysis were used to regress the population structure shown in Fig. 6.21 to a theoretical original distribution at hour 5 (i.e. prior to the timecourse). This distribution is best described in relation to the beginning of S phase, with t = 0 being the point where cells enter S phase.

The first of these methods involved studying the most probabilistically varied distribution of the data by regressing each timepoint separately rather than considering the progression of one timepoint to the next. This method is a summing of probabilities. If x cells were in early S phase at hour 15, these cells could have begun early S phase as early as hour 9 and as late as hour 15, given that early S phase is 7 hours long (as determined in Section 6.1). In the original distribution, each hour between 9 and 15 would therefore have an average probability of being the starting hour for x/7 cells. If the populations of each timepoint are regressed in this manner (additionally using the durations for mid and late S phase also determined in Section 6.1), a distribution is formed that centres slightly prior to S phase entry. (see Fig. 6.24).

![Fig 6.24 – The population structure estimated at Hour 5 using a probability based method with the population at each of the timecourse hours being regressed individually.](image)

The spread of the data is extreme, with a standard deviation of 4.1 hours around a mean of t = -1.6. However, this wide spread is due to the treatment of each population as a separate entity. If this method is tested on an ideal timecourse result (as shown in Fig. 6.25a), which would occur due to a single peaked population concentrated in a single hour, the distribution is still wide (with a standard deviation of 3.4 hours around a mean of t = -1.8 hours) due to use of probabilities (see Fig. 6.25b). Despite this spread, it would appear that the test
population most likely stems from a single peaked distribution centred on \( t = -1.6 \), the majority of which would begin S phase in hour 7.

![Graph showing population distribution across hours after nocodazole release](image)

**Fig. 6.25a)** - The ideal timecourse results that would be expected for a population that centred around \( t = -2 \), aimed to begin S phase in Hour 7 of the timecourse. The resulting timecourse is extrapolated using the S phase durations determined in section 6.1 (a 7 hour early S phase, a 2.14 mid S phase and a 1.55 hour late S phase).

![Graph showing population distribution across cell cycle hours](image)

**Fig 6.25b)** - The population at hour 5 that is derived from the theoretical results shown in a) using the probability based method. Whilst the peak is correctly centred, there is a wide degree of spread.

In the second method of analysis, the populations were instead regressed through determining the potential path through S phase taken by cells beginning in each hour interval, and quantifying the likelihood that that path could have been taken. Rather than treating each
population as independent data, this method sought to recreate the dependency that is inherent to the progression of the population. When this analysis was performed on the ideal timecourse data, an underlying population centred entirely within \( t = -2 \) (so as to begin S phase in hour 7) was uncovered (see Fig. 6.26).

Fig 6.26 – The population at Hour 5 that is derived from the theoretical results shown in 6.25a) using a path-based probability method.

However, this form of analysis requires more information than the previous method, as cells that would begin S phase near the end of the timecourse or end S phase near the start of the timecourse have less time points to negotiate. To prevent the dominance of shorter paths, the likelihoods of cells progressing were weighted according to the number of timepoints the path was judged by. The likelihood was therefore weighted through multiplying by (number of timepoints covered/11), with 11 being the most timepoints a cell could occupy (i.e. its entire S phase lay in the timecourse). The results of this regression method are shown in Fig. 6.27.

Fig 6.27 – The population at Hour 5 that is derived from the experimental results shown in Fig. 6.21 using a path-based probability method.
The analysis again shows a peak of cells that would begin S phase in hour 7. The spread of data was now more concentrated to the left of this peak, indicating that after the initial mass entry of cells into S phase at hour 7, frequent entries will also occur at subsequent timepoints. With regards to the timecourse experiment, this would indicate that values taken early in the timecourse are less likely to be influenced by noise than later values. However, due to the method of weighting required for this analysis, the exact ratio of the peak versus the noise cannot be certain.

As the distribution had been shown to have a single peak value, analysis of the timecourse could occur through a study of the differences between one fork rate profile and the next. The major difference between timepoint t and timepoint t+1 would be that the peak of population would have progressed one hour further into their S phase (see Fig. 6.28). Once the fork rate profiles were scaled to the same overall population size and smoothed to mitigate the effects of noise, the profile from t could then be subtracted from t+1 to give an indication of the changes that occurred. As this often resulted in a cluster of additional values, a median of the remaining peaks was then taken. This process was repeated for a range of windows of smoothing size and an average taken. The fork rate profile generated through this method is shown in Fig. 6.29. The wider range of fork rate values explains more of the variation seen in the data, and the distribution also has a mean fork rate of 1.38 kbp/min, which is similar to the average fork rates of 1.49 kbp/min for each of the control populations tested in section 6.2.
The same method of analysis was then begun on the origin density profiles. However, upon smoothing of the data, it became apparent that the origin profiles shared a similar median with the timepoints estimated to represent S phase all lying between 64 and 88 kbp (see Fig. 6.30). As the median values were fairly consistent and do not follow any pattern that could be related to the fork rate profile, from the data available it can be concluded that the density of origins does not change during a normal S phase. As fork rates occur over a 0.5-2.2 kbp/min range, it is possible to calculate from the calibration curves developed in section 6.2 that the origin density values would need to range from 30 kbp to >100 kbp if they were to compensate for the altered fork rates. In the entire dataset, only 5% of origin lengths are less

Fig. 6.28 – An illustration of the analytical method used to extract the most recent changes in the fork profile. Profile shows the distribution of the population at each hour, with the dotted line marking the beginning of S phase. To extract the fork rate for each time point, the cumulative previous hour(s) data is subtracted from the current replication fork profile to reveal the leading edge of the data. Hatched area shows the cumulative data to be subtracted.

Fig. 6.29 – The fork distribution created after taking the median of the data remaining after the processing step shown in Fig. 6.28.
than 40kb, hence the adaptation of origin density towards fork rates within a normal S phase would appear not to occur, at least to the extremes seen during replicative stress.

Fig. 6.30 – Median replicon lengths from the 10 hours of the timecourse that approximate to the majority of cells S phase. The profile shows little relation to the processed or unprocessed replication fork data. n = 14-65 for each timepoint.

Fig. 6.31 – The replicon length profiles of the 10 approximated hours of S phase. Each colour indicates the profile of a different hour. The profiles show considerable overlap, with variation most likely due to the small n in some of the datasets.
6.4.4 Conclusions

A number of conclusions that are relevant to future modelling can be drawn from this data. Firstly, the fork rates over an MRC5 cell’s S phase vary in a similar way to those observed in HeLa cells by Takebayashi et al (2005). The generated distribution has a suitable mean fork rate for MRC5 cells as determined in Section 6.2 and it explains much of the variation observed in the timecourse experiment.

The profiles of origin densities were seen to vary little over the course of S phase, with the majority of the replicon counts from each hour interval showing considerable overlap. This would indicate that fork rates are likely to be changing speed in response to a factor innate to the DNA, such as the chromatin state, rather than due to origin spacing. The slight changes in origin density appear to be due to natural variation in the data, likely due to low n values in some datasets. The variation that is seen is not significant enough to demonstrate a balance to the changes in fork rates. The observed distribution of replicon lengths is insufficiently varied when compared to the calibration curves generated in Section 6.2 and does not follow the pattern that would be expected in response to the replication fork profile.

It is noted however that the median of the origin densities is lower than that observed in HeLa cells (Jackson & Pombo, 1998), which could potentially lead to altered structuring of replicon clusters in MRC5 nuclei, such as a greater number of replicons per cluster. This effect may be cell type specific.
6.5 Perspective

The analysis of the S phase across cell populations has provided an estimation of the S phase duration for the experimentally studied cell line, in addition to quantifying the length of early, mid and late S phase. These estimated lengths are similar to accepted values for other tested cell lines. However, the figures are unlikely to be accurate enough to determine whether the slight observed differences in duration are due to the experimental counts or due to variation between cell lines.

The experimental population’s reaction to varied fork rates and origin densities has also been quantified, demonstrating the bi-directional relationship that occurs as part of a monitoring system to maintain consistent activity at the replicon level. Whilst this relationship is demonstrated to function during replicative stress, experiments indicate that little or no alteration in origin density occurs due to fork rate changes during a normal S phase.

Using the synchronisation method developed in Section 6.3, a population of cells was partially synchronised and its S phase characterised in terms of progression, replication fork rates and origin densities. Both of the profiles generated in Section 6.4 can provide data for further modelling. The experiment has generated a fork rate profile specific to the MRC5 cell line which spans the full extent of S phase. From the origin density profiles, an average distribution can also be taken, which will give a distribution of origin densities which is also specific to MRC5 cells. Further analysis should note the potential for a greater number of replicons per cluster than HeLa cells, or that there will be a greater number of clusters if the profile of replicons per cluster is maintained.
7.0 - Genomic Linear Models of the Mammalian Cell S Phase

With the refined datasets created in Chapter 6, a final model of the mammalian cell S phase was developed to integrate the novel data with the previous modelling developments from Chapters 4 and 5. The aims of this final iteration of the model are to indicate areas requiring further study and to determine what conclusions can be derived from modelling given more accurate parameterisation.

7.1. Modelling Method

The model created in Chapter 5 demonstrated how the use of replicon cluster firing probabilities could circumvent the lack of data concerning replication factory progression. The firing probability mechanism represents a method of summarising the unknown data required to program activation into a single set of variables that can be manipulated and tested as required. Given the success of this technique in the model developed in Chapter 5, the same mechanism was used in the creation of the model documented in this chapter.

Given some of the concerns raised in Chapter 4 regarding the treatment of chromosomes as a separate entities (see Sections 4.3.3 ii) and vii), the model was designed to simulate an entire diploid genome, featuring a set of 46 chromosomes. This process was facilitated by the removal of the replication factory as an active component of the model. Previously, it was assumed that the diffusion of factories would require a 3 dimensional model of the nucleus to be created, which would in turn require data that is currently unavailable in order to parameterise it correctly. However, cluster firing probabilities can be applied across the genome without concern as to the structure. Chromosomes were therefore treated as separate entities for the purpose of neighbour-driven activation, but results (e.g. DNA
replicated over time) were drawn from the genome as a whole. A diploid genome of 46 chromosomes (X,Y) was specifically chosen to best simulate the MRC5 genome for which the variable fork rate distribution and the origin distribution had been determined in Section 6.4.

In order to successfully guide overall replication levels, a profile of DNA replication over time was produced from the TimEX-seq profile of embryonic stem cells (Desprat et al., 2009, data processed by Pedro Olivares-Chauvet). The use of experimental data was deemed to be more accurate than the estimated of 10% completion per hour, again aimed at creating more biologically realistic results.

One aims of the model will be to study the possible dynamics of cluster activation that can accurately create the experimentally determined profile of DNA replication. Given the potential for a range of solutions to this profile, the outputs of the model will also be compared to the following datasets:

2) *De Novo* cluster firing and single-direction and bi-directional activation of neighbouring clusters (as described in Section 5.1).

*Modelling Inputs*

In addition to the 46 chromosome framework, the model was built with the following input parameters:

1) Experimentally determined fork rates for MRC5 cells as derived in Section 6.4. The distribution has been smoothed (for reasons illustrated in section 5.1), giving the replication fork profile shown in Fig. 7.1.
ii) Experimentally determined origin distribution for MRC5 cells. This profile is derived from the replicon lengths measured in the Section 6.4 which have been combined to give a general distribution on the assumption that the distribution does not vary over time.

iii) A profile of DNA replicated over time, summed over the entire genome from the Desprat et al (2009) human erythroid TimEX-seq dataset (see Fig. 7.2) (values extracted by Pedro Olivares-Chauvet).

Fig. 7.1 – Experimentally determined replication fork profile for MRC5 cell line. Blue lines shows the raw data. Red line shows the smoothed profile used in simulations.

Fig. 7.2 – Cumulative DNA replicated over time for human erythroid cells, extracted from the TimEX-seq dataset.
As replicon lengths are to be drawn from a distribution derived for the observed MRC5 replicon distribution, the data required approximation to a mathematical function. The data failed a Jarque-Bera test of normality, and appeared better suited to a log-normal distribution (mu = 4.3043, standard deviation = 0.3978) (see Fig. 7.3) with a minimum replicon length of 20 kbp.

![Fig. 7.3 - Fitting of distributions to the experimental dataset. Stars indicate the fitted distributions.](image)

Left – A normal distribution fitted to data.

Right – A log normal distribution fitted to the data.
7.2 Are varying activation probabilities required?

Using the method described in Section 7.1, a genome-wide analysis of DNA replication could now be performed, with each iteration of the model simulating DNA replication over the 46 human chromosomes of a diploid cell.

Using this framework, it was possible to test the previous assumption that timely DNA replication cannot occur through constant replicon cluster activation probabilities. The model created in Chapter 5 had indicated this was likely on the level of individual chromosomes, but the interaction of replication across the genome may have lead to different observations. With a firing probability of 0.1/60 for each cluster (averaging 10% of the genome every hour), the resulting duration of DNA replication had too long a tail, extending far beyond a 10 hour S phase (see Fig. 7.4a), and also did not match the DNA replication profile indicated by the TimEX-seq dataset (see Fig. 7.4b).

Fig 7.4a – Cumulative DNA replicated over time. Red line shows TimEX-seq Data and blue line shows simulated output. Data averaged over 100 iterations. Due to the large number of clusters active, the maximum standard deviation from the mean over the timecourse is only 27,000 kbp/min.

Fig 7.4b – DNA replicated over time. Red line shows TimEX-seq Data and blue line shows simulated output. Data averaged over 100 iterations.
These results indicate that more complex activation dynamics also occur during genome-scale operation of DNA replication, and that the required dynamics are not just an artefact of the limited scale of the model in Chapter 5. Varying activation probabilities will therefore once again be considered.

7.3 Calculation of cluster activation dynamics

Analysis of the model generated in Chapter 6 demonstrated that the programmed dynamics of the model could lead to the generation of very varied behaviour with regard to the observed outputs. Whilst these outputs and the criteria that judged them gave a loose conclusion as to the preferred type of dynamics, the potential parameter space that was allowed remained relatively large. The large range of cluster firing probability profiles and neighbour activation settings led to the creation of many sets of possible activation dynamics, whilst the choice of a specific format for the probability profiles (a sigmoidal curve) may have limited the model. The model described in this chapter was designed to avoid the potential for behavioural assumptions, whilst also reducing the parameter space as far as possible.

The use of the TimEX-seq data set to limit DNA replication removed the ambiguity in judgement, as it is a definite descriptor of expected dynamics. To allow as close a simulation as possible, the two modes of cluster activation (neighbour activation and de novo) required calculation with as few assumptions as possible, and ideally in concert with each other. For this model, neighbour activation was defined as the probability (tested from 0 to 1) that a completed cluster could activate either or both of its neighbouring clusters, with each neighbour being tested independently.

The model was therefore developed in two iterative parts. During the first, the number of cluster activations required to complete enough DNA replication to match the TimEX-seq profile was calculated. This criterion was judged at each minute timepoint. However, in the
case of neighbour activation, completed clusters had the potential to activate their neighbours before the firing of *de novo* clusters, allowing the neighbour influence to be incorporated as the firing probability profile was generated.

The resulting firing probability profile and the specified neighbour activation setting were used to parameterise the secondary iterations of the model. The firing profile was first smoothed, as the calculation of the firing probabilities in the first step led to an exactly optimised profile with very specific peaks of activation that appear biologically unrealistic (see Fig. 7.5). Observations of the dynamics were then drawn from this second set of simulations. This process had the benefit of allowing the model to attempt to fit the expected data (the TimEX-seq profile) without forcing the dynamics to obey a specific mechanism (such as the dynamics observed in the model generated in Chapter 4).

Fig. 7.5 – A cluster firing probability profile determined by the first step of the model using a neighbour activation probability of 0.2. Red line shows the original profile. Blue line shows the smoothed profile to be used in the second step of the model.
7.4 Results

A number of specific questions were used to direct the testing of the model and provide biologically relevant conclusions. It should be noted that the standard deviation of DNA replicated over time is not included in the following graphs; as demonstrated in Fig 7.4a, the large number of active clusters across a genome leads to relatively uniform replication through each iteration.

i) How does neighbour activation affect cluster replication dynamics?

The model was tested over a range of neighbour activation probabilities from 0 to 1. At each of these settings a profile of cluster activation was generated and then tested through the second set of iterations. The use of 100% neighbour activation forced the model to over-activate clusters at several time points. This was due to de novo clusters being activated to meet the overall replication profile which would then activate their neighbours with too high a frequency (see Fig. 7.6).

Fig. 7.6 – DNA replicated over time. Red line shows TimEX-seq Data and blue line shows simulated output using a neighbour activation probability of 1. Data averaged over 100 iterations.

However, if lower probabilities of neighbour activation were used, fitting of the simulated DNA output to the TimEX-seq data could be improved, with a minimum deviation seen at a probability of 0.7  (see Fig. 7.6a and b). Nearest neighbour activation combined with fitted de
*de novo* cluster activation probabilities therefore provides a method with which the TimEX-seq data can be more closely simulated.

**Fig. 7.6a** – The average deviation from the expected DNA replication output. The difference is calculated at each timepoint and then averaged across the timecourse. Deviation is then averaged over 100 iterations at each neighbour activation probability.

b) DNA replicated per minute at probabilities of 0.6, 0.7 and 0.8 neighbour activation. Timecourses are averaged over 100 iterations. Red line indicates expected DNA replicated per minute.
ii) Do fitted activation settings give realistic dynamics?

Whilst the firing probabilities and neighbour activation settings can be fitted to the expected DNA output, the dynamics that result may not be supported by other experimental data. To test this, the modes of cluster activation were measured across the timecourses for each neighbour activation setting. Firstly, the observed single to dual sided activations ratio was recorded. Cluster activation was judged to have been a neighbour activated if it occurred within one hour of its neighbour’s completion. If a cluster has only one neighbour satisfying this criterion, a single sided activation event was recorded. If both its neighbours fired within this timeframe, a dual sided activation event was recorded. If a cluster was activated independently of either of these mechanisms, it was classified as *de novo* activation. Observations were only recorded after the first hour of activations, with percentages drawn only from the clusters activated after this period, as the criteria dynamics are representative of an ongoing S phase rather than the initial period. The results of this study are shown in Fig. 7.7.

![Graph showing the behaviour of activation over a range of neighbour activation setting from 0 to 1. Blue line indicates the percentage of activations which are *de novo*. Red line indicates the ratio of single-sided activation to dual-sided activation. Results are averaged over 100 iterations.](image)

Interestingly, the percentage of *de novo* activations is not 100% even without any possibility of neighbour instigated activation. This is due to the appearance of activation events being coupled with neighbours due to chance *de novo* activations. This observation could have
important consequences for analysis of *in vivo* experiments, as it would indicate a base level of associated firing even if there is no genetic continuity between cluster activation.

At the neighbour activation probabilities that best fit the observed replication over time data, the dynamics of cluster activation appear to display a likeness to the experimental data. Probability settings of 0.5 to 1 give a single to dual activation ratio of ~5:1. The *de novo* activation remains high however, with 30% of clusters being *de novo* activations at the optimal probability setting of 0.7.

Another discrepancy in the simulated data is shown when compared to the origin activation profile of Goldar *et al* (2009). Simulations were run with a neighbour activation probability of 0.7, monitoring the origin firing over chromosome 6 during 1,000 simulated S phases. The averaged proportion of remaining origins firing per minute over the timecourse is shown in Fig. 7.8. This profile is comparable to the experimentally observed origin firing profile, which indicated the proportion of origins firing from those that remained. As occurred in the model developed in Chapter 5, the profile appears to peak too late in S phase, and each timepoint varied by an average of +/- 20.0% from the experimentally determined profile.

![Fig. 7.8 – The proportion remaining origins firing per minute at a neighbour activation probability of 0.7. Results are averaged over 1000 iterations.](image-url)
iii) **Could variable origin densities give a more appropriate origin firing profile?**

A potential solution to the discrepancy in the activation profiles is that the assumption of constant origin density is false. If origin density were to respond to fork rates, one would expect an increase in origins fired during the timepoints of slower fork rates, which would coincide with the peaks seen in the expected origin firing profile.

To test this hypothesis, simulations were run as in Section 7.4 ii) but with the allowance for altered origin densities. This was achieved through altering the length of replicons within clusters as they were activated so that each replicon was proportioned to the fork speed at the time of firing. To achieve this with minimal alteration of cluster completion times, an average time required to replicate the cluster was calculated from the original cluster. The total length of the cluster was maintained, whilst the fork rate at the time of firing was used to determine how many replicons of equal length would best fit the desired completion time. This method could potentially lead to an alteration in the number of origins, as would be expected if the density is proportional to fork rates. The overall time required for clusters was generally lower as a result of more efficient clustering of replicons, and thus a lower neighbour activation probability of 0.6 gave optimal replication dynamics when compared to the TimEX-seq data. Using this method, origin firing was again monitored for chromosome 6 during 1,000 simulated S phases. The averaged proportion of remaining origins firing per minute over the timecourse is shown in Fig. 7.9.
The alteration of replicon clustering leads to a slight increase of variation between the experimental profile and the simulated profile however, with values from the simulated profile being an average of +/- 21.2% from the experimental profile. Despite a close fit of DNA replication to the TimEX-seq data in both sets of simulations, it would appear that the model cannot simulate the origin firing profile, even with the addition of favourable alterations of origin clustering. Given the experimental data’s stipulation of origin firing as origins fired per time unit per unit length of unreplicated DNA, a higher peak of origins firing would have to be witnessed during mid S phase. However, the level of firing required to would cause DNA replication to greatly exceed the amount indicated by the TimEX-seq dataset.

7.5 Conclusions

The use of the variable fork rate profiles and origin densities derived from the same cell line has provided a close approximation of both DNA replication over time (through comparison to the TimEX-seq dataset) and to S phase activation dynamics. From these simulations, activation of clusters throughout S phase would be predicted to operate with ~70% neighbour activation, with the remainder of activations required being de novo. At these high levels of neighbour activation, such a mechanism could occur through the certain activation of clusters...
if they are contacted by oncoming forks. Under this hypothesis, 30% of outgoing forks would either stall (unlikely, given the potential for DNA damage) or would traverse inter-cluster regions of DNA. The extra time required before activation due to the buffer-zone of DNA would give an increased likelihood of the next-in-line cluster having already been activated as part of the de novo activated group of clusters. If 30% of extruded forks were delayed in this manner, an overall neighbour activation of 70% would occur despite the mechanism operating on a certainty of firing once contact was made.

The model supports the general idea of cluster firing probabilities increasing over time (as shown in Fig. 7.5), with the profile giving rise to accurate and timely replication of DNA. A slight increase in firing probabilities during mid S phase is also probable, so as to counter the reduced fork rates (which may occur due to chromatin-transition zones). Alternatively, this increased probability may occur as the result of the increased activity of additional cyclin activation complexes such as Cyclin A2-Cdk1 (Katsuno et al., 2009). The increased activity of this complex would lead to the introduction of a new target group of clusters whilst early clusters were also still being completed. This overlap could potentially lead to an increase in cluster activation activity which is then reduced upon the completion of all early clusters.

The model is unable to simulate the origin firing profile of Goldar et al (2009) however. This appears to be due to the peak in origin firing density at mid S phase in the profile being unnecessary to meet the quota of DNA replication at that time. Increased firing probabilities towards late S phase (and hence a slightly higher number of origins fired per length of DNA) lead to the timely completion of replication which matches the TimEX-seq profile. This discrepancy remains inexplicable by the model.

Also noted during the simulations were the occurrences of apparent neighbour activations despite the prohibition of direct neighbour influence. A basal level of around 20% of
activations could simply appear to be neighbour related when they are in fact de novo. However, this level is too low to mitigate the conclusions of Maya-Mendoza et al (2010).

7.6 Perspective

It has been demonstrated that accurate inputs can lead to a comprehensive model of the mammalian cell S phase, allowing the exploration of DNA replication and different theoretical concepts. The model cannot explain all of the experimentally observed phenomena, implying that there may be further intricacies to the system that need to be incorporated.

The model also illustrates the importance of using data derived from similar, or ideally, the same, cell lines. The reduction in replicon lengths in MRC5 cell compared to HeLa cells leads to a variation in the predicted structure of replicon clusters. Further studies could therefore measure the grouping of replicons within clusters in the MRC5 cell line to determine the universality of the structure of clusters.

The model supports both the concept of next-in-line activation and the increase of cluster firing probabilities as S phase progresses.
8.0 Conclusions

This study has explored a range of concepts within the subject area of the mammalian cell S phase using a variety of in vivo experiments and in vitro modelling. From the results of these experiments, a number of conclusions can be drawn which are relevant to future study within the field.

i) Cluster activation methods can be predicted through modelling approaches

The models developed in Chapters 5 and 7 each determine that close approximations of the experimental observations can be simulated through a combination of de novo and neighbour activation mechanisms. Although in comparison the earlier model predicted a lower level of neighbour activation, its governing rules were less biologically realistic than the final model. The models agree that neighbour activation is required to prevent unsuitable levels of de novo cluster activation. The models also both indicate that if the probability of neighbour activation is too high, cluster activation after the first round of cluster completion can lead to unrealistically high levels of DNA replication. The final model therefore places the optimal probability of neighbour activation at ~0.7.

This level of neighbour activation could occur as a result of two potential mechanisms. Firstly, 30% of forks extruded from completed clusters may stall during their progression to subsequent clusters. This process, repeated very frequently over the S phase, could lead to potential DNA damage however.

The second potential mechanism is one that the model has not yet accounted for due to the lack of experimental data which would be required so as to correctly parameterise the formation of chromatin structures. In all models developed in this study, each simulated cluster is perfectly adjacent to the next, with no DNA required as a linker. This allows extruded forks to progress instantly from one cluster to the next. However, clusters may be
linked by longer lengths of DNA than merely the two most external replicons, with DNA formations perhaps up to 1 to several Mbp in length if chromatin structures include Giant Loops (Cremer et al, 2006). There are also transition zones of DNA, such as those identified between R- and G-banded regions, which feature low origin density and may therefore require replication by an encroaching fork from their periphery. These two varieties of ‘buffer’ zones between clusters have not been imposed on the current cluster-based framework of the model. An estimate by Farkash-Amar & Simon (2010) is that 10% of the mouse genome may occupy regions demonstrating uniform fork direction. This DNA would require distribution around the existing model as either of the two types of buffer zones.

With consideration of these zones, forks could potentially be allowed a 100% chance of activating any cluster they enter. This certainty of activation would be mitigated due to the requirement that extruded replication forks must traverse the inter-cluster buffer zones before activating subsequent clusters.

The progression of forks into these zones could result in three possible outcomes. If the region is sufficiently short, the extruded fork would activate the next-in-line cluster within a close enough time frame as to be identified as a neighbour activation. Alternatively, as zone length increases, so too does the probability that the next-in-line cluster will undergo de novo activation. Finally, the fork may traverse the zone and activate the cluster, but the time required would give the appearance that the activation is not coupled.

Interestingly, if the sum of these possibilities is a 30% reduction of neighbour activations, this would create a mixture of dynamics seen between the neighbour activation probabilities of 0.7 and 1. DNA replication output would occur at a similar rate to that seen in the 0.7 probability profile (i.e. the best fit) whilst the actual number of de novo activations that occur would move towards the 10% value seen with a probability of 1. The single:dual cluster activation ratio remains at ~5:1 in either profile, hence all the cluster activation dynamics would be satisfied.
ii) Identification of S phase replication patterns

This study has illustrated the difficulties in automating the identification of S phase replication patterns. One particular source of variation that has been highlighted is the disparity in patterns generated when labelling nuclei using different staining methods.

The calibration of the algorithm developed in Section 6.1 involved the use of a test set consisting of nuclei stained using three different methods (BrdU, Biotin and EdU) combined with a number of different secondary antibodies and chemical stains. This test set was designed in this way so as to contain examples of all of the potential images that would require processing and identification.

Whilst the use of a mixed test set could allow the generation of separation techniques that allow correct segregation of patterns in all possible datasets, this could only occur if the variation between the patterns remained greater than the variation resulting from the use of different labelling techniques. A more successful method may involve the separation of the datasets, with images being supplied with metadata describing the labelling technique. The algorithm could then be programmed to alter rules and separation boundaries dependent on the expected patterns exhibited due to the staining method. This separation of the datasets may also help overcome the variation caused by differing intensities which were more pronounced when comparing differentially labelled images.

Another process that could allow easier comparison of nuclei is the standardisation of nuclei shape and size. This process was attempted during testing of the algorithm but was found to morph nuclei in a disproportionate manner if the nuclei were not roughly spherical. However, such nuclei could be removed through the imposition of a stringent cut-off for nuclear circularity during the particle analysis step (i.e. the step when the nuclei are initially identified). This would be an appropriate filter if the proportion of non-spherical nuclei is the same at all stages of the S phase. Such nuclei could then be removed without biasing the results. Nuclei with standard shape and size could be more easily compared to an identification template for each of the patterns under the different staining methods.
iii) Replication fork rates vary significantly during the mammalian cell S phase

The experiments conducted in Section 6.4 demonstrated that replication forks progress at different speeds throughout S phase of the MRC5 cell line. The profile generated is similar to that generated for Hela cells by Takebayashi et al (2005). The maximum and minimum fork rates vary greatly around the mean fork rate, although a reason for this has not yet been identified. However, the identification of a similar profile with closely matching fork rates at comparable timepoints in two cell lines likely indicates that the profile is common to mammalian cells. This may imply that the causal factor is also shared or has a common distribution in different cell lines.

iv) Origin Density in the MRC5 cell line is greater than that of the HeLa cell line

The density of origins with the MRC5 cell line appears to be different from the HeLa cell line, with a lower median inter-origin distance. This may imply either that replicon clusters in MRC5 nuclei contain a higher number of replicons on average than with Hela cells, or that there are a greater number of replicon clusters within each MRC5 nucleus. This occurrence may represent a general alteration in the structuring of chromatin between the two cell lines. This could have arisen as a result of adaption of transformed Hela cells to long term growth in culture. Alternatively, the two cell lines are also differentiated by both their transcription and their origins in relation to their original tissue types.

v) Origin density remains constant over the MRC5 S phase

The distribution of origins within MRC5 cells has been experimentally demonstrated to remain constant over time. Whilst variation does occur in the length of replicons, this appears to be evenly distributed around a fairly constant median. This would amount to replicon clusters being relatively uniform in loop size (although overall cluster length may vary) which would
be expected if their formation occurs as a result of traits common to chromatin such as persistence length. This would however imply that the varied origin density of HeLa cells would occur as a result of an alteration the general nature of chromatin. A more dynamic process (such as influence of previous S phases) may therefore be responsible, which could be more easily varied between cell lines whilst remaining consistent across a population within the same cell line.

The alternative hypothesis of origins having varied distributions over S phase has been tested and has been shown to give no better fit to the experimental data. There would therefore appear to be little evidence within this study, either experimentally or inferred from simulations, that origin density varies over S phase in the MRC5 cell line.

Additionally, it was noted that the origin distribution which was compiled from the individual profiles featured a long tail, with small numbers of very long replicons. These replicons may be representative of replication through the transition zones, and are not present in each individual profile. However, this may be a result of either the sampling n being too small to identify extreme cases or the difficulty in detecting distant origins pairs due to breaks in fibres that occur during the DNA fibre analysis protocol.

### vi) Association of replication timing to R- and G-banding patterns

The study conducted in Chapter 5 illustrated the inaccuracies in the approximation of replication timing to R- and G-banding. A number of suggestions were made as to why this could have occurred. One set of explanations concerns the nature of the giemsa banded dataset. The data obtained from the UCSC genome browser may be of insufficient resolution to allow comparison of replication timing, with smaller domains being classified as parts of larger domains despite having a different replication period within S phase. The data may also not be able to account for the transient and dynamic nature of chromatin, with areas displaying a gradient of properties between different epistates that are not expressed within the dataset.
An alternative to this branch of explanations is that replication timing is not as closely associated with R-/G-banding as has previously been assumed. The loose correlation may be due to association of replication to another factor that is in turn associated with R-/G-banding, such as gene expression.
9.0 Perspectives

Whilst this study has provided information on a number of areas within the general study of the mammalian cell S phase, some of the conclusions made are limited and require further exploration. Other areas have proved to be interesting research concepts, yet have not been fully investigated within the scope of this study. This section will evaluate the conclusions made and make suggestions for further research.

i) Can modelling accurately describe the mammalian cell S phase?

This study has demonstrated the progression of modelling of the mammalian cell S phase from independent simulations of single chromosomes to a model of the genome that displays biologically realistic replication dynamics. The process has identified the inconsistencies of approximating replication timing with R- and G-banding, and has allowed the exploration of both next-in-line activation (Sporbert et al., 2002) and the concept of increasing origin firing probabilities (Hyrien et al., 2003; Rhind, 2006).

The iterative steps of the model development have also allowed the identification of the key parameters that require accurate parameterisation in order to give realistic dynamics. The creation of novel datasets with which to populate the model has been demonstrated to yield insights into the behaviour of the system and predictions about the mechanisms which operate. Whilst some information (such as the origin firing profile) has not yet been reconciled with any of the modelling steps, modelling can provide a platform from which to test why these differences occur. Such events can indicate areas where either the modelling theory is incorrect (which can be refined as demonstrated during the study), more appropriate data is required or the underlying biological mechanisms require review.

There are a number of possible explanations for the discrepancies between the results of the model and the temporal profile of origin activation developed by Goldar et al. (2008). The origin activation profile was created through identification of the peaks in the replication
timing profile of Woodfine et al. (2004). Each peak indicated the firing of at least one origin, which allowed the creation of a minimal profile of origin activation.

The models developed in this study have two features that may prevent them from adhering to the estimated profile. Firstly, in order to ensure an S phase under 10 hours in duration, a high probability of cluster activation is calculated during the later time points in the model. This forces high levels of origin activation in the last hour of S phase. However, modelling of the *S. pombe* genome by Lygeros et al. (2008) led to the suggestion that replication can continue into G2, which would remove this requirement for a late burst of origin activation. The relative peak of origin firing would thus move closer to mid S phase.

Alternatively, the models created in this study also monitor the activation of all origins that are fired during the timecourse. In the final model, clusters are automatically activated through next-in-line activation rather than undergoing passive replication, and hence they and the internal origins are recorded as having been activated. A cluster activated through the next-in-line method may not necessarily create a peak in the replication timing profile of Woodfine et al. (2004) however, either due to the resolution of the data or the lack of texture created in the profile by short replicon cluster. Passively replicated origins (or even passively replicated replicon clusters if they are not automatically activated) would not be visible at all. These variations in resolution and mode of scoring likely create the imbalances seen between the experimental and the simulated data.

Further dispute in the measuring of origin firing performed by the models is likely to occur as a result of the assumption of synchronised firing of a small number of perfectly efficient origins in each replicon cluster. This mechanism may be an over-simplification of a system that is carefully calibrated to prevent DNA damage from fork stalling within clusters. Intra-cluster origin firing in a format applicable to the model has been simulated by Blow & Ge (2009). It was found that inefficient origins could protect against replication fork failure whilst requiring less resources than a multitude of active origins. The simulations also demonstrated that replication could occur from a number of efficient, but not synchronised, comparatively ‘early-firing’ origins. One could therefore predict replicon clusters to contain a gradient of origin efficiencies, which the current model does not display. The use of such
detailed intra-cluster activation of origins could also make completion more efficient given
the possibility (especially with random replicon clustering) of a high inter-efficient-origin
distance. Given the likelihood of 1-2 inefficient origins within this region, cluster completion
would be more uniform due to the possibility of inefficient origins filling the ‘gaps’. Should
the model also feature the potential for fork stalling, dormant origins may be essential in
ensuring a successful S phase.

The final model developed in Chapter 7 created a profile of increasing firing probabilities for
clusters as S phase progressed, which occurred in concert with the neighbour activation
mechanism. This profile represented an integration of the all effects other than neighbour
activation which could result in varied firing probabilities. It is therefore essentially the sum
of the two separate activation profiles developed in the model of Chapter 5, which
represented the activation probabilities of clusters in R- and G-bands. Whilst the distinction
between R- and G-banded clusters was removed for the final model, other models choose to
instead explore the differentiation of activation through early and late, hence removing the
ambiguity caused by the assumption of replication be accurately tied to banding patterns.

Such modelling techniques (and others) have been recently reviewed by Rhind et al (2010).
Two main questions arise from the culmination of these techniques; why do origins (and in
the context of this study, replicon clusters) have varying activation probabilities, and why do
these probabilities change over time?

The models developed in chapters 4 and 5 each demonstrate one of the potential causes of
overall variation, i.e. banding patterns. Rhind et al summarise two causes- firstly variable
chromatin structure (linked to the banding patterns) and secondly the potential for non-
uniform MCM distribution. It is reasoned that areas of increased licensing are more likely to
form origins than areas of lower licensing given a particular probability of firing which is
consistent between clusters. Such a disparity in licensing may occur if one were to compare
euchromatic regions of the genome to transition zones. This second reason would again call in
to question the use of a single distribution of origin density as used in the models developed
in this study.
The reasons for varying firing probabilities over time are divided into three main groups in the review. First is the availability of a limited activator complex. Whilst the levels would remain constant, the reduction in total DNA to replicate over time would lead to a relative increase in activation factor compared with the remaining origins/clusters, resulting in increased activation probabilities. Along a similar line is the recycling of polymerase complexes. Either of these two reasons would be comparative to the overall limitation of DNA replication as used in the models developed in Chapters 5 and 7.

The third group follows a dynamic not attempted in this study, which is the increased availability of an activation factor over time. This could occur as the result of the gradual import of a factor from the cytoplasm to the nucleus (Rhind et al., 2010). Such a method would compound the relative increase that would occur from the previous two mechanisms.

In addition to the groups indicated in the review, influence of cyclin complexes (as approximated in Chapter 5) may also be responsible for alterations in firing probabilities over time due to control they exert over the system. Temporally separate groups of origins would require variation in the structure of origins, specifying that their activation must occur only during times of activity of a particular cyclin complex. This method could interplay with the previous mechanisms to provide an external control as a result of the checkpoint and DNA damage pathways.

The model developed in Chapter 7 creates an activation profile describing what the sum of these influences on activation may look like. Subsequent modelling may therefore attempt to break this profile down into its constituent elements, such as the imposition of early and late firing differentiation, activation properties intrinsic to origin structure or the influence of control complexes. Such an analysis would require a detailed understanding of the processes that combine to create the carefully orchestrated behaviours that the system demonstrates.

Whilst the models developed within this study do not yet accurately represent the entirety of the mammalian S phase, the increasing complexity of the models, as further understanding has been incorporated, has led the closer simulation of biological observations. The value of information extracted from a model is dependent on the assumptions and the data that form
the basis of the model. With further development, the model system will likely simulate mammalian cell S phase increasingly accurately, but observations must be evaluated with regard to the inputs of the model. It should also be impressed that as a tool for understanding and exploring the S phase, an entirely accurate model may not be required.

**ii) How are replicons clustered?**

Two sets of parameters have been used to describe the formation of replicon clusters throughout the models developed in this study. A draw from the distribution of replicons per cluster indicated the overall structure, and replicons were then sampled independently and randomly from a distribution of lengths. However, it may be appropriate for another parameter to be added which describes the similarity between replicons within a cluster. This parameter would represent an important influence on the structure of replicon clusters as it would determine the efficiency of replicon cluster completion. Selection of origin spacing may be relatively flexible in the case of licensing groups of pre-RCs (Lebofsky et al, 2006), with the selection of origins from the groups being a function of a variety of influencing parameters.

The congruency parameter would be influenced towards similar replicon lengths due to the standard persistence length of chromatin combined with the selection for increased efficiency with regard to more rapid replicon cluster completion. A cluster of similarly sized replicons will almost certainly achieve a higher level of DNA replication over time than a less coordinated structure. This would in turn be balanced towards incongruent clustering by factors such as the influence of transcription and variation in the progress of DNA replication over the previous S phases (due to the influence of one generation’s S phase on the next (Courbet et al, 2008)). Interference between active origins and pre-RCs could influence the parameter in either direction depending on the nature of the interaction over given distances of DNA. One would expect a standard level of inhibition from active origins, giving rise to
consistent inter-origin distances, but the distance that this interaction functions over (i.e. intra-licensing-group or inter-licensing-group) may mitigate the effects.

The distribution used for the drawing of replicons per cluster may also require additional consideration. The models described in this project used a distribution observed in HeLa cells (Jackson & Pombo, 1998) as no distribution is available for other cell types, despite the use of other MRC5 derived datasets in the final model. Given that the replicons are on average shorter in the MRC5 cell line than in the HeLa cell line, use of this distribution means that the model has incorporated the assumption that replicon clusters are more numerous in the MRC5 cells line and are on average smaller in size.

As replicon size has been shown to vary between cell types, testing of other parameters such as the number of replicons per cluster would be a sensible precaution. Such a process would either lead to more accurate modelling or greater confidence in the existing model.

It is also important to note that the clustering of replicons within this study has only involved the simulation of what are likely the most efficient origins of replication. This is due to the sampling of lengths from an observed distribution of replicons, which would result only from origins which have fired. Given the possible role of less efficient origins in closing of gaps (Blow & Ge, 2009), this additional level of detail could also be imposed on the formation of clusters, with the seeding of additional origins (perhaps derived from caffeine treated cells to cause maximal origin firing) over the observed distributions.

Given these arguments and observations, an experimentally determined set of grouped replicon lengths, perhaps observed through biotin labelling, would likely be a very useful resource for further modelling. Even if the dataset only described coupled replicons, the implications of the data could still be useful in estimating the likely structure of replicon clusters.
iii) **What causes variation in fork rates?**

The experiments measuring replication fork rates in MRC5 cells demonstrated that fork rates do vary considerably during a normal S phase. Previous experiments by Conti et al (2007) in human keratinocytes found a linear correlation between fork rates and inter-origin distances. However, from the data gathered in MRC5 cells, replicon length clearly did not fall to the level required to compensate the lower fork rates. Variation in origin density may potentially affect fork rates, but it does not appear to maintain the constant level of replication demonstrated through the experiments with caffeine and aphidicolin (see Section 6.2). One may therefore consider alternative, or perhaps additional, reasons for the reduction in fork rates.

The recognised correlation between replication and transcription (Gilbert, 2002; Woodfine et al, 2004; Jeon et al, 2005) could provide a number of explanations for this observed gap between fork rates and origin density. Given that the replication of DNA containing housekeeping genes occurs in early S phase, it is probable that replication machinery will at some point encounter transcription machinery also attached to the DNA. Negotiation between the two is likely to at least cause slowing of the replication complexes. Alternatively, the areas of DNA replicated during early S phase are those most likely to be vital for survival of the cell. Selection may therefore have driven for slower replication in these areas in order to avoid potentially lethal replicative errors.

During characterisation of the profile from HeLa cells, Takebayashi et al (2005) described areas of slow fork progression as replication slow zones, which would be replicated with enough co-ordination during S phase to lower the average fork rates to the levels measured. These zones were approximated to R/G-band boundaries, with the reduction in fork rate aiding the temporal separation of R and G-banded DNA. These transition zones are also areas of low origin density, hence fork rates may slow whilst obstacles to replication are dealt with, as fork collapse would be less likely to be rescued by the firing of additional origins.
There are a number of potential reasons for the fork rate variation, at least some of which may be shared between cell lines. The exact reason currently remains unknown however, and may be the subject of further study.

iv) How much does the S phase vary in duration?

This study illustrated some of the difficulties observed when synchronising populations of cells to make S phase measurement. In addition to the initial challenges faced when grouping the cells around a time point in a synchronised fashion, this study is likely to have been affected by the progression of cells at different rates through S phase (as demonstrated in other studies such as Okeefe et al (1992) and Jackson (1995)). When conducting experiments measuring S phase dynamics across cell populations, the natural variation inherent to the system creates additional difficulties during analysis. This is compounded by the consideration that synchronicity is increasingly lost over time. The gradual loss of synchronicity may have been observable in the fork rate profiles shown in Fig. 6.22, where it appears that early timepoints have larger and less numerous peaks than later timepoints. Alternatively the additional noise may have been due to the increasingly complex S phase population at each subsequent timepoint as more cells outside the synchronisation peak entered S phase.

Given that this degree of variation of S phase progression timing occurs within a single population of one cell line, there is also then the question of whether there is greater variation still between different human cell lines.

If DNA replication timing is linked to the nature of chromatin as a result of gene expression, one would expect a predictable amount of variation between cell lines dependent on the level of differentiation. This would be exhibited as variation in the replication patterns, different temporal profiles of DNA replication over time or through changes in the overall duration of S phase. Each cell line would be expected to exhibit replication patterns and temporal profiles specific to its chromatin state. However, within populations of each cell line there would then be the potential for variation away from this profile, either due to the potentially stochastic nature of DNA replication or due to variations in the state of chromatin.
between cells. This variation from the averaged profile could be expected to be of a constant degree when compared between cell populations given a fairly consistent program of gene expression within each population. However, the variation between cell lines is likely to be less easily predicted, with the potential for a different average durations and different temporal profiles for each.

This variation not only implies the need for large n values when sampling populations, but also the need for a level of tolerance during modelling simulations. Recreation of the S phase replication patterns, for example, is likely to display many sub-optimal patterns if the process is performed stochastically. Interpretation of the results will reply partially on judging then expected degree of variation, and then judging how much of the remaining variation is due to inconsistencies between the model and the in vivo subject. An accurate description of the S phase duration is unlikely to be described beyond population-wide probabilities without single cell data however.

v) How varied are S phase parameters in different cell lines?

This study identified similarities between the variable forks rate profiles derived from two human cells lines during S phase. However, other factors, such as the density of origins, were shown to vary, which is likely to also result in variation in either the structure of replicon clusters or the number of clusters that are formed across the genome. The differentiation exhibited between cell types is of interest to future modelling efforts, as it creates boundaries to inferences between different cell lines. The reasons for this variation are also of great interest. Similarities would result from shared traits between cell lines which, in the case of the fork rate profiles, create similar replication dynamics. Differences between S phase parameters would in turn result from causal factors which are varied between cell lines; hence key S phase influences could be identified through comparison of the characteristics which differentiate cell lines.

A link between replication and transcription has already been implicated as being a causal factor in orchestrating S phase dynamics. Section iii) suggested as to why variation within the
fork rate profile of a single cell line could occur as a result of transcription. Given the correlation between transcription and replication, transcription could in turn be responsible for some of the similarities exhibited between cell lines. Although gene expression will be altered between cell lines, it will result in the common alteration of certain areas, such as an open chromatin configuration of regions containing housekeeping genes. These areas are likely to share chromatin epistates, and during S phase would be the sites of similar dynamics between replication and transcription machinery. The observed fork rate profiles would therefore display a similar shape, independent of the locations of DNA replication over time. However, without precise determination of the causal factor for fork rate variation, comparison between cell lines remains speculative.

Explanation of the differences between S phase in different cell lines is equally difficult. Variations in average origin density could occur for a number of reasons. It may be that a slightly higher number of MCMs are loaded on to DNA in the MRC5 cell line, resulting in shorter replicon lengths on average. The estimation of this is difficult due to the possibility of licensing groups however (Lebofsky et al, 2006), with uncertainty as to how many MCs are loaded and how they co-ordinate firing (see Fig. 2.1). The interference in activation of pre-RCs by active origins through the ATR pathway may regulate origin density to a predetermined level despite the loading of additional MCMs.

Alternatively there may be alterations in either replicative or transcriptional machinery that favour slightly different chromatin structures. Experiments by Courbet et al (2008) studying fork rates and origin densities demonstrate that the dynamics of replication of one S phase in turn influences the next. In this way, slight alterations could be passed on through cell line generations. Given the potential for both genetic drift and changes in nuclear architecture in cell lines cultivated in the laboratory environment for many years, accumulated variation could easily lead to very different S phase dynamics.
vi) **How can we progress into three-dimensional modelling of the S phase?**

One of the aims of this study was to attempt to move into three-dimensional modelling of the mammalian cell S phase in order to allow comparison of whole nuclei observations to simulated nuclei. In order to achieve this aim, a number of obstacles must first be overcome.

A key issue is the lack of accurate information describing the location of DNA within the nucleus. Whilst replicon clusters can be randomly allocated, or formations of DNA created through modelled approximations such as fractal clustering (Lieberman-Aiden et al., 2009), the attributes governing the properties of DNA (such as gene content) must also be correctly distributed. Whilst general assumptions can be made, such as euchromatic DNA being concentrated towards the centre of the nuclei, the complexity of creating a realistic framework of chromatin without the availability of nuclear coordinates makes the process largely speculative.

However, such a model, with an approximation of DNA distribution, structuring and properties using rules that govern its most likely conformations, could be used to study the potential factors which relate to DNA replication. With a bias towards euchromatin at the centre of the nuclei, one would expect a rough simulation of the S phase patterns to occur if the genome was programmed to undergo DNA replication in a fashion similar to the linear models in this study. Allowing preferential activation of euchromatic areas in early S phase would result in replication across the centre of the nucleus followed by the periphery as progression to mid and late S phase ensued. However, simulation of the patterns seen in late S phase would likely require a more complex distribution of DNA. The areas which are replicated last are distributed across the nucleus in dense clumps, and their formation would likely require additional parameterisation so as to accurately simulate heterochromatin.

Such a model could be potentially tested through parameterising the DNA with a description of the open/closed state of the chromatin. Activation of clusters could be programmed to occur in a similar manner to the model described in Chapter 5, with firing probabilities altered according to the state of the chromatin. Alterations of this measure of chromatin compactness could then be compared to an *in vivo* experiment where trichostatin A is used to
de-condense the DNA. Changes in the replication patterns are then likely to occur both in vivo and in the simulated nucleus. The similarity of these changes will depend on the accuracy of the distribution of DNA and the chromatin compactness descriptor in addition to the accuracy of the correlation between DNA replication and chromatin compactness.

Three-dimensional modelling could also make use of the availability of accurate three-dimensional scans of nuclei. With the introduction of super-resolution microscopy (Baddeley et al, 2010; Koberna et al (2005)), this level of scanning could be used over multiple nuclei to create probability maps as to where replication is likely to occur in particular stages of the S phase. Perturbations to chromatin structure could then be used to study how these maps would be altered.

Given the likelihood of further information describing the structure of the nucleus, the development of modelling into three dimensions will likely be possible as further data is discovered. More biologically appropriate theory is supported by advances in microscopy which can provide increasingly accurate parameterisation for models. In combination, these factors will allow the creation of a model directly comparable to the experimentally observed patterns of replication across the nucleus.

**vii) Can the identification of nuclear replication patterns be automated?**

This study has indicated a number of sources of variation in nucleus-wide DNA staining that lead to difficulty in the automated identification of S phase replication patterns. These sources of variation could likely be overcome through study of each potential variation source (such as label used) in an independent manner. The sorting algorithm could then be programmed to use specific identification rules according to the metadata supplied with the images.

In addition to the alteration of rules, more detailed scanning could also aid the process. The use of three-dimensional scans would provide additional relevant information, such as providing extra datapoints with which to estimate the nuclear periphery highlighted in the
case of mid S phase cells. Multiple depth scans could also prevent the misidentification of patterns due to the particular plane of scan that has been captured. During scans of nuclei, it was noted in many cases that the manual identification of cells was influenced by the consideration of multiple planes. A mid S phase nucleus’ top- and bottom-most planes can be mistaken for an early S phase nuclei, and it is only through the consideration of the characteristic staining pattern in the central planes that a correct identification can be made.

The creation of three-dimensional probability maps of replication (as discussed in vi)) could also be useful in the identification of patterns, as three-dimensional scans could be compared to the maps and the best fit solution identified.

An alternative to the comparison of three-dimensional scans to replication probability maps would be possible provided an approximation of the distribution of DNA across the nucleus. With the availability of such a guide to the nuclei, the observed staining patterns could be fitted to a map of the euchromatic/heterochromatic compartments across the nucleus and a ratio of replication within the two used to determine the current stage of the nucleus in S phase. The localisation of replication within dense areas of heterochromatin across the nucleus during late S phase is again likely to complicate the analysis however.

Given further investigation, the process of automated replication pattern identification is likely to be possible; however, the correct parameterisation will require careful calibration for different staining and cell types, and would be best performed as part of a dedicated software package.
9.1 Study Summary

This investigation began with an extremely simplified and naive model of DNA replication and has demonstrated the process of iterative testing and parameterisation in order to develop the model through a number of stages. This process has driven experimental study and has generated many further questions that may be of interest to the field. Combined with continued developments in both experimental and analytical tools, this continued process will likely lead to more accurate and explicit model of the mammalian cell S phase, with greater predictive power. This combination of modelling and experimental work can therefore be used to both explore the system and test our assumptions in order that a more complete understanding is gained.
10.0 References


Jares P, Blow J: Xenopus cdc7 function is dependent on licensing but not on XORC, XCdc6, or CDK activity and is required for XCdc45 loading, Genes Dev 2000, 14: 1528-1540.


Maya-Mendoza A, Petermann E, Gillespie D, Caldecott K, Jackson D. Chk1 regulates the density of active replication origins during the vertebrate S phase. EMBO J. 2007, 26:2719-31.


11.0 Appendix

Referred from Section 4.1 - Binned data for replicons per cluster (approximated from Jackson & Pombo (1998)).

Replicons per Cluster Distribution

<table>
<thead>
<tr>
<th>Number of Replicons</th>
<th>Cumulative Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>2</td>
<td>0.29</td>
</tr>
<tr>
<td>3</td>
<td>0.48</td>
</tr>
<tr>
<td>4</td>
<td>0.71</td>
</tr>
<tr>
<td>5</td>
<td>0.83</td>
</tr>
<tr>
<td>6</td>
<td>0.90</td>
</tr>
<tr>
<td>7</td>
<td>0.94</td>
</tr>
<tr>
<td>8</td>
<td>0.97</td>
</tr>
<tr>
<td>9</td>
<td>0.98</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
</tr>
</tbody>
</table>