5. Results 3
Using Genomic Microarrays to assess Replication Timing in a Human Cell line and correlation with sequence features

5.1: Introduction

DNA microarrays have been used to assess replication timing in the yeast *Saccharomyces cerevisiae* (Raghuraman, Winzeler et al. 2001) and in *Drosophila melanogaster* (Schubeler, Scalzo et al. 2002). Assessment of replication involves timing measuring the difference in copy number in an S phase population, between sequences that replicate early in S phase (and are therefore present in a high copy number) and those that replicate late in S phase. This Chapter describes how genomic arrays have been used for the first time to assess human replication timing. In these experiments differentially labelled S and G1 phase DNA is co-hybridised to the arrays described below. Loci reporting a ratio close to 2:1 are early replicating; conversely loci reporting a ratio close to 1:1 are late replicating.

Initially the hybridisations were carried out on an array representing the entire genome with clones at a 1Mb resolution. Section 5.2 describes the replication timing profiles obtained for each chromosome at this resolution. Correlations between the replication timing and sequence features of the genome at this resolution are also shown.

The S:G1 hybridisations were also performed on arrays constructed at tile path resolution. A large, a medium and a small chromosome are examined at this resolution, using tiling path arrays for chromosomes 1, 6 and 22. Analysis of replication timing at this higher resolution enables more accurate mapping of zones in which a transition in replication time occurs. The tile path arrays and the correlations with sequence features at this level are described in section 5.3.

Finally, the replication timing of a small region of chromosome 22 is assessed at a very high resolution. PCR products were used to cover a 4.5Mb region of chromosome 22 at 10Kb resolution. A region of 200Kb was analysed at an even

higher resolution using overlapping 500bp PCR products. This array is described in 5.4.

The replication timing method was substantiated by comparison with alternative ways of assessment of replication timing. The replication timing of 11q was compared with published data for this region and the results are described in section 5.5.1. Assessment of the replication timing of clones within chromosome 22, described by the array as early, mid or late replicating was performed by real-time PCR and the results correlated with the array data. This is reported in section 5.5.2.

During DNA replication, copy number increases from one to two at each locus. This should occur early in S phase for early replicating loci, but in later fractions of S phase for later replicating loci. This can be detected by the hybridisation of fractions of S phase to the array. S phase nuclei were sorted into four fractions and co-hybridised with G1 phase DNA on the 22 tile path arrays. This allowed comparison of the array analysis with assay of replication timing using quantitative PCR, in which S phase is conventionally divided into four fractions and the fraction in which a predetermined loci replicates is resolved. This is described in section 5.7.

5.2: Assessment of Replication Timing on the 1Mb array

5.2.1: Obtaining the Average Replication Timing of Individual Chromosomes.

The S:G1 ratios and standard deviation of each locus in 4 replicate 1Mb genome wide arrays were calculated and the average co-efficient of variation of all loci in the four replicates was found to be 5.5%.

The ratios for all the loci contained within the same chromosome were averaged and are shown in Figure 5.1 and Table 5.1.

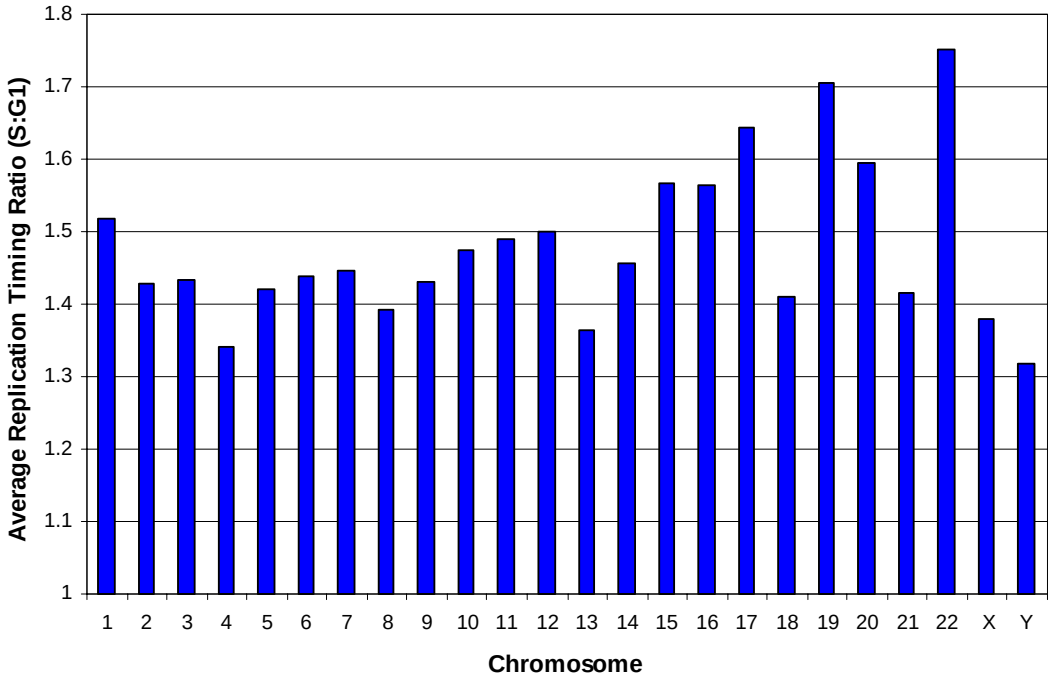


Figure 5.1: The average replication times of all 24 chromosomes.

Table 5.1: The average replication times of all 24 chromosomes (Early-late)

Chromosome	**Mean Replication Timing Ratio**
22	1.75
19	1.72
17	1.64
20	1.60
15	1.57
16	1.56
1	1.52
12	1.50
11	1.49
10	1.49
14	1.46
7	1.45
6	1.44
9	1.44
3	1.43
2	1.43
5	1.42
18	1.42
21	1.42
8	1.39
X	1.38
13	1.36
4	1.34
Y	1.32

The average replication timing for each chromosome was used to normalise the individual chromosome tiling path arrays.

5.2.2: Correlating Chromosomal Replication Timing with Sequence Features of the Genome.

The chromosome wide individual sequence features were plotted against the replication timing of each chromosome as shown in Figure 5.2.

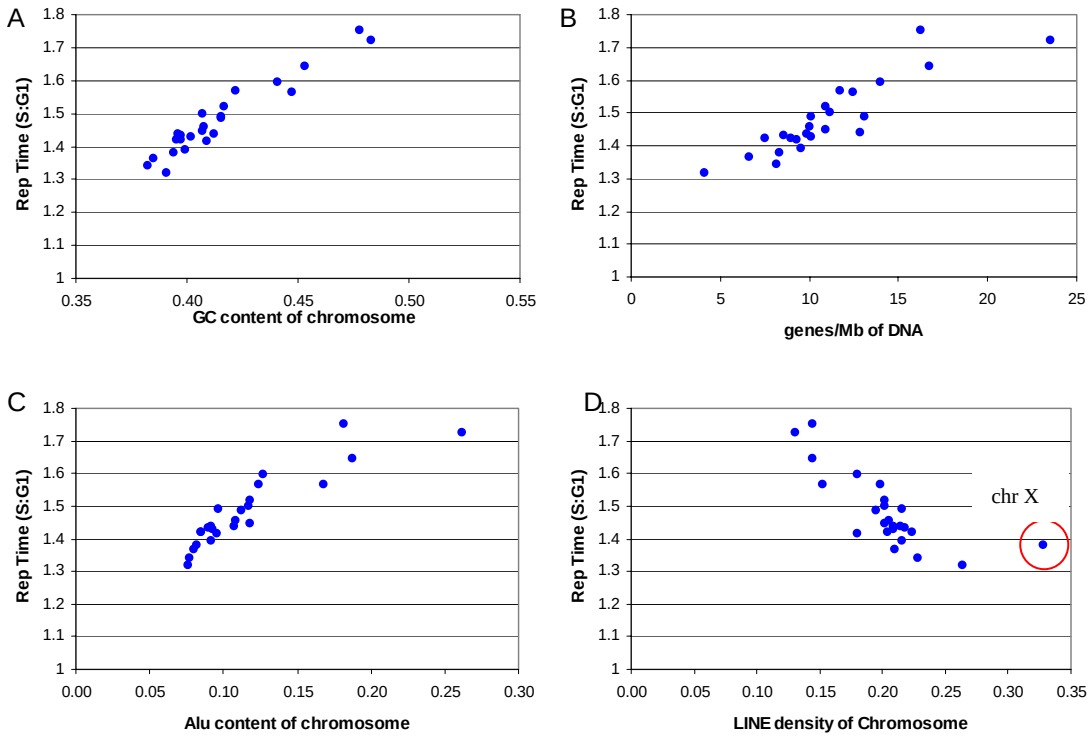


Figure 5.2: Correlation between Replication Timing and Sequence Features of the Genome. A: Correlation with GC content. B: Correlation with Gene Density. C: Correlation with *Alu* repeat content. D: Correlation with LINE density.

Linear regression was then performed on these data to assess the correlation between replication timing and sequence features. The linear regression statistics are shown in Table 5.2.

Table 5.2: Linear regression performed between replication timing and genome statistics.

Genome Feature	**Regression Coefficient (x)**	**Intercept (y)**	**Correlation Coefficient (r)**
GC Content	0.04	0.17	0.96
Gene Density	0.02	1.20	0.89
***Alu* Repeat Content**	0.02	1.21	0.9
LINE Repeat Content	-0.02	1.91	0.72

Significant correlations were found for all sequence features although the best correlation with the S:G1 ratio was found with GC content. A very strong positive correlation was seen with *Alu* repeat content and a strong positive correlation was observed with gene density. A strong negative correlation was observed between

LINE repeat content and replication timing; however the correlation coefficient is not as strong as those observed with other features. Definitions of the strength of correlations is found in Appendix 13.

5.2.3: Assessing Replication Timing at a 1 Mb resolution.

Data from the same array experiments could also be used to assess replication timing at a 1Mb resolution. Each clone on the array was positioned according to the NCBI 31 assembly of the human genome as represented in Ensembl. The replication timing ratio was then plotted against position on the chromosome to produce a replication timing profile for each chromosome (Appendix 6). Example replication timing profiles for two chromosomes are shown in Figure 5.3

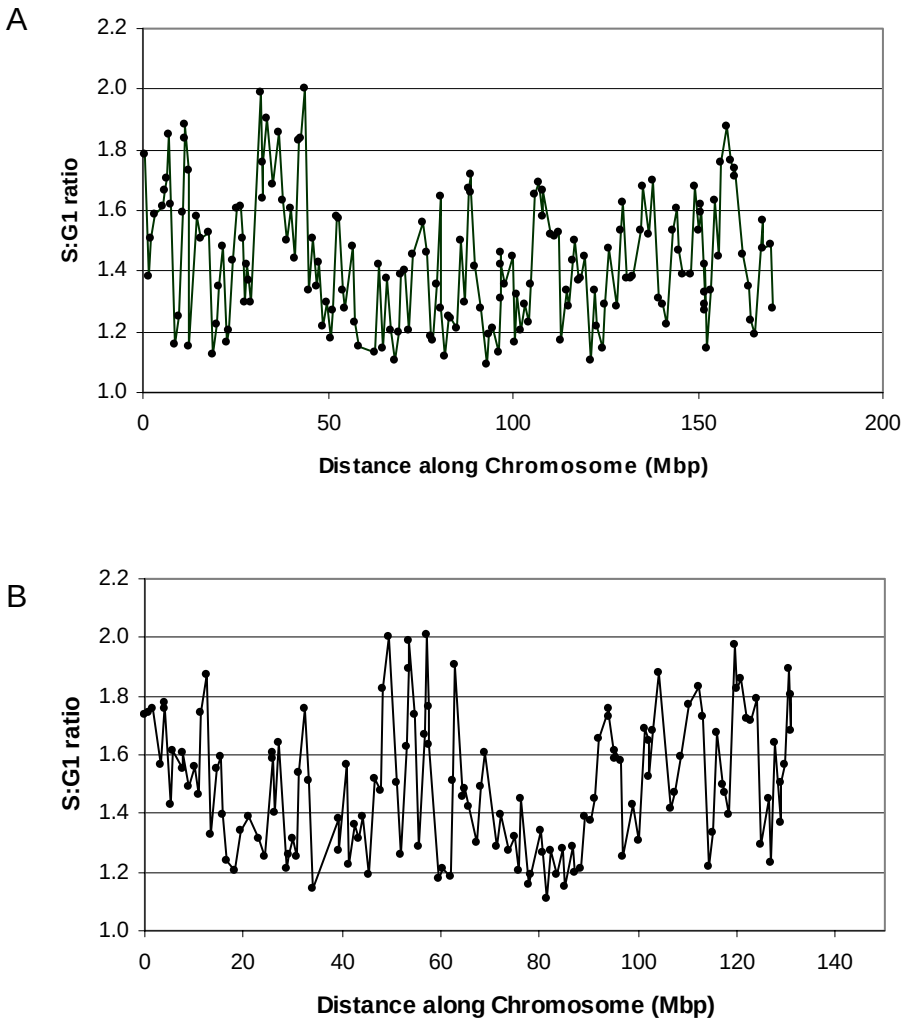


Figure 5.3: Replication timing profiles of; A: chromosomes 6 and B: chromosome 12

Correlations between replication timing and sequence features were also performed with the whole genome sampled at a 1Mb resolution. The sequence features used for the correlation were for just the clone sequence represented on the array, not the whole 1Mb window represented by that clone. The correlations are shown in Figure 5.4.

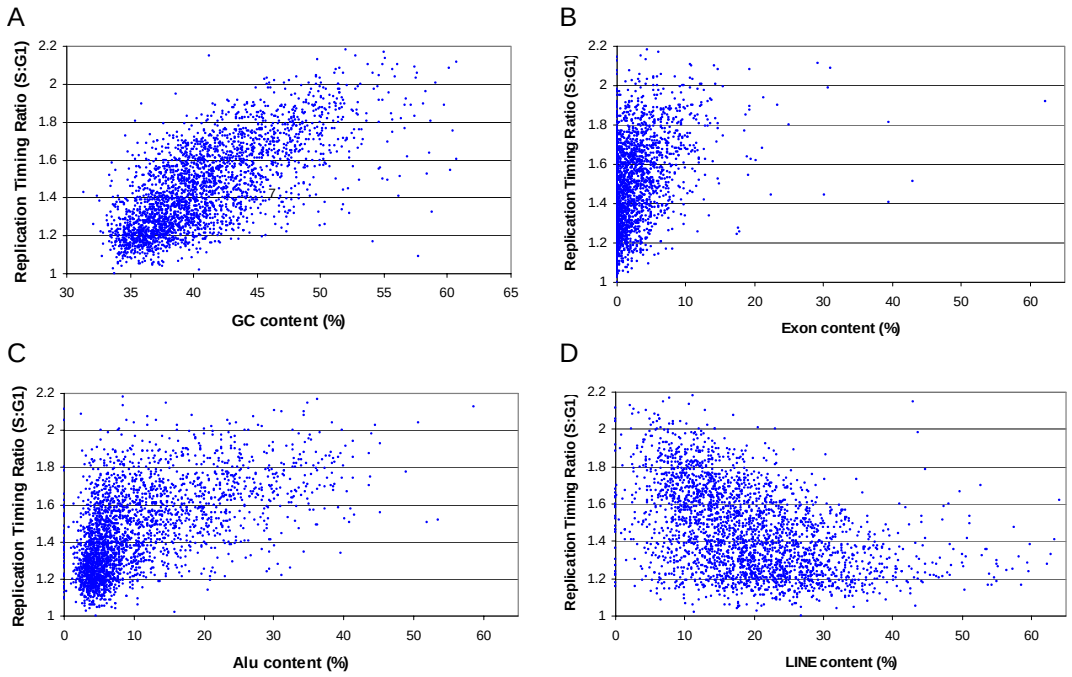


Figure 5.4: Correlation between Replication Timing and Sequence Features of the Genome. A: Correlation with GC content. B: Correlation with Exon Density. C: Correlation with *Alu* repeat content. D: Correlation with LINE density.

Linear regression was then performed on this data to assess the correlation between replication timing and sequence features. The linear regression statistics are shown in Table 5.3.

Table 5.3: Linear regression performed between replication timing and genome statistics at a 1 Mb resolution.

Genome Feature	**Regression Coefficient (x)**	**Intercept (y)**	**Correlation Coefficient (r)**
GC Content	0.032	0.12	0.70
Gene Density	0.002	1.37	0.35
Exon Density	0.023	1.40	0.42
***Alu* Repeat Content**	0.014	1.30	0.56
LINE Repeat Content	-0.008	1.62	0.40

The best correlation was again seen with GC content. Positive correlations were also observed with *Alu* repeat content, exon density and gene density. A negative correlation was observed with LINE repeat density.

5.3: Assessment of Replication Timing at Tile path Resolution

Replication timing was assayed using tiling path arrays for chromosomes 1, 6 and 22. These arrays were constructed using DNA extracted from sequencing clones. This ensured the data obtained provided complete coverage for chromosome 22q, chromosome 6 and chromosome 1.

5.3.1: The Replication Timing of Chromosome 22.

Replication Timing was assayed on the chromosome 22q tiling path array. The S:G1 hybridisation was carried out on four separate arrays to ensure reproducibility of the data. The average co-efficient of variation between the four replicates was found to be 5.5%. The replication timing of chromosome 22q is shown in Figure 5.5.

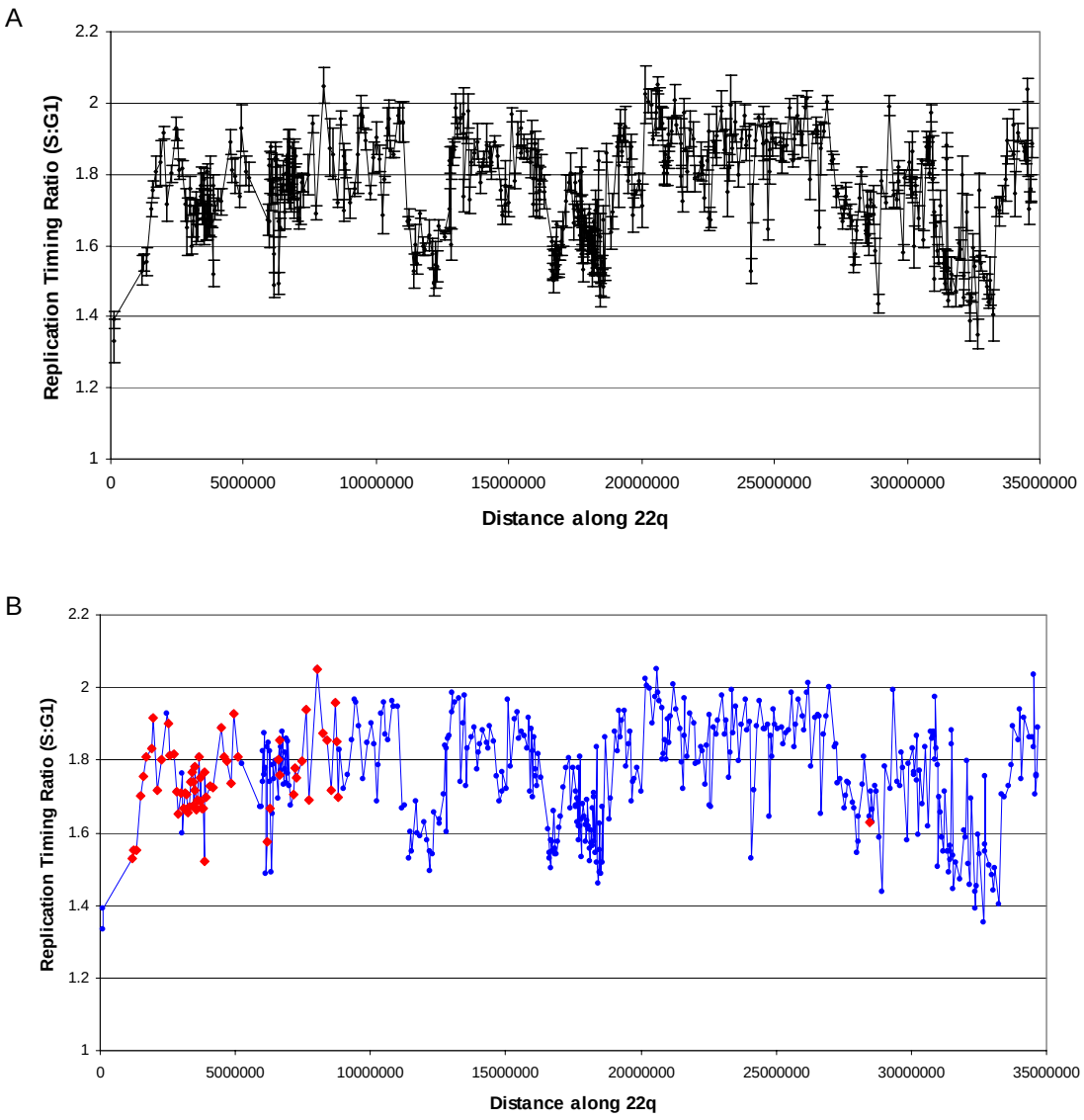


Figure 5.5: Replication Timing profile of Chromosome 22. A: Replication Timing profile of chromosome 22 showing the standard error of each loci (n=4) on the Y error bars. B: Replication Timing profile of chromosome 22 with clones containing significant amounts of segmental duplication highlighted in red (Buckley, Mantripragada et al. 2002).

Chromosome 22 contains considerable amounts of segmental duplication at 22q11. This has a significant impact on the ability of clones on the array to accurately report locus-specific copy number changes as sequences with homology to these regions will cross hybridise with the DNA on the array. These clones are highlighted on Figure 5.5B. As a result the replication timing of the arrays reported by these regions will be inaccurate and a composite of the replication timing of all the regions sharing the

homology. As a result these clones are not included in further analyses assessing the correlation of replication timing with other sequence features. Replication timing was plotted against sequence features of the genome such as gene density, GC content and common repeat content. These correlations are shown in Figure 5.6.

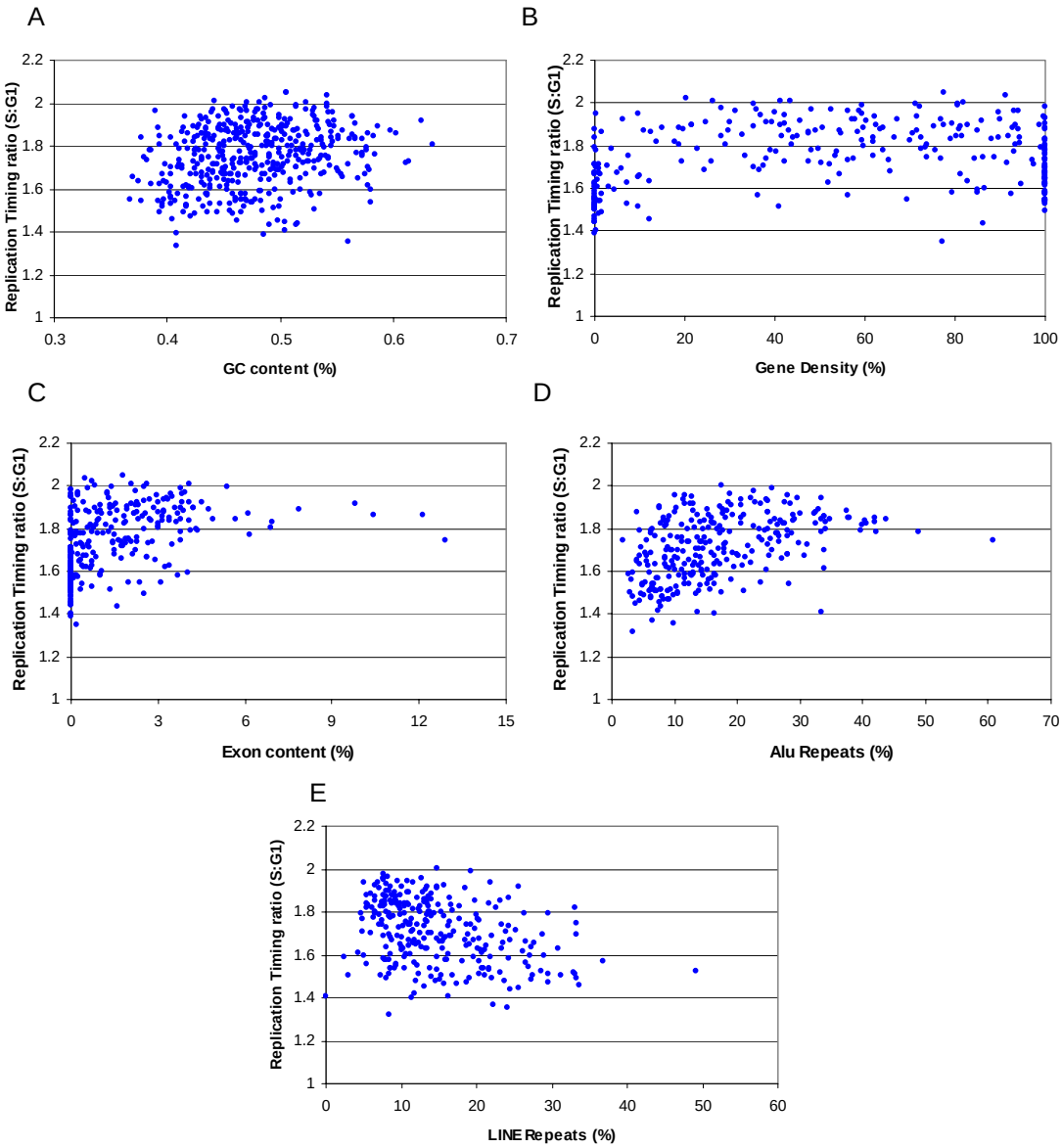


Figure 5.6: Correlation between replication timing and other genome features on the 22 tile path array. A: Correlation with GC content. B: Correlation with gene density. C: Correlation with exon density. D: Correlation with *Alu* repeat content. E: Correlation with LINE repeat content.

Linear regression was then performed on this data to assess the correlation between replication timing and sequence features. The linear regression statistics are shown in Table 5.4.

Table 5.4: Linear regression performed between replication timing and genome statistics at a 78Kb resolution on the 22 tile path array.

Genome Feature	**Regression Coefficient (x)**	**Intercept (y)**	**Correlation Coefficient (r)**
GC Content	0.63	1.45	0.22
Gene Density	0.001	1.71	0.19
Exon density	0.03	1.70	0.39
***Alu* Repeat Content**	0.007	1.59	0.50
LINE Repeat Content	-0.007	1.81	0.34

The genome features were plotted against chromosome position (Figure 5.7) together with replication timing to allow comparison of areas where the correlation between replication timing and other features is strong, and regions where correlation was weak.

Figure 5.7 shows that replication timing follows the general patterns of change in the various different genome features. It can be seen that regions such 30-34 Mb along 22q show a very poor correlation with GC content, yet show very good correlations with exon density and *Alu* repeat content.

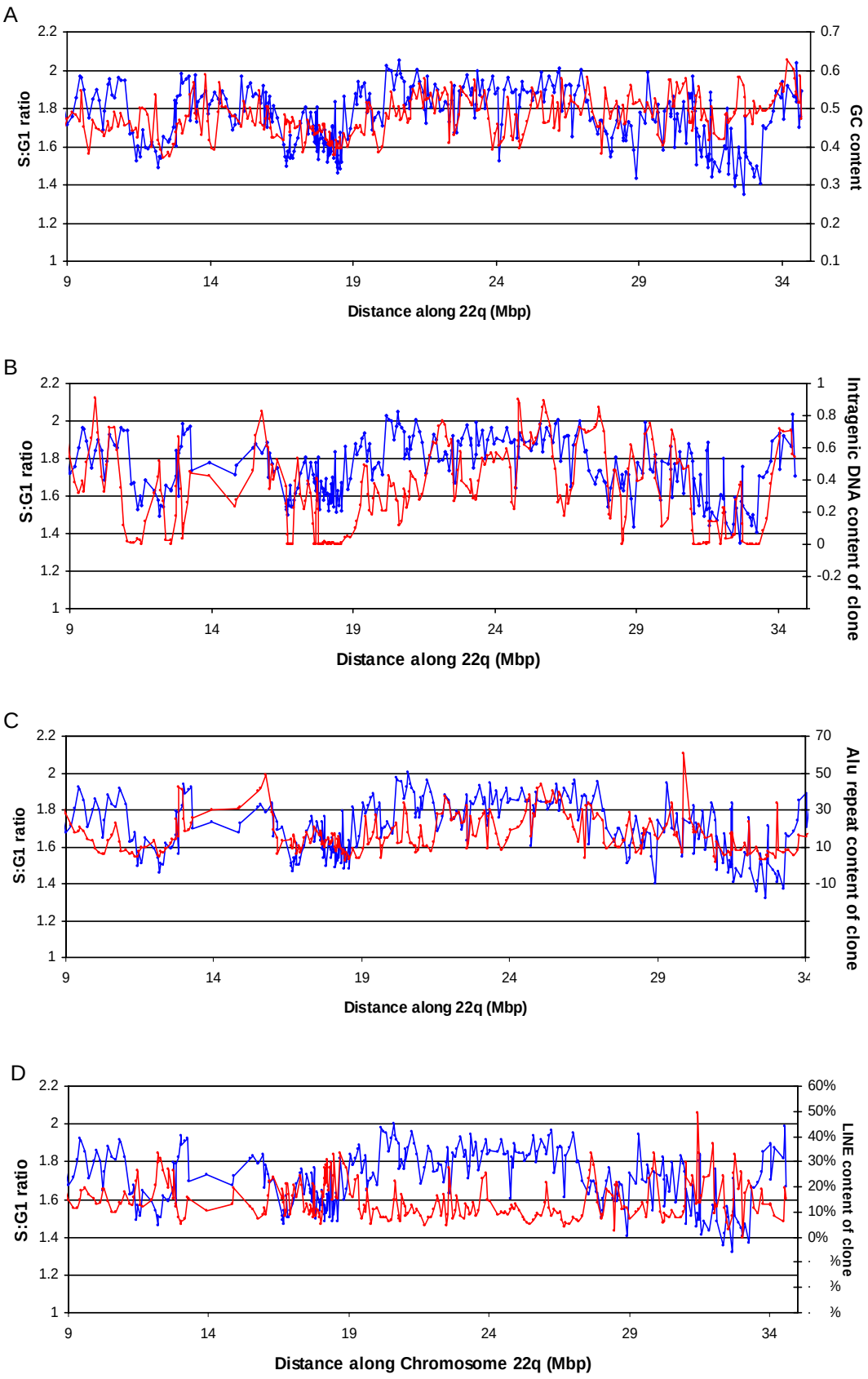


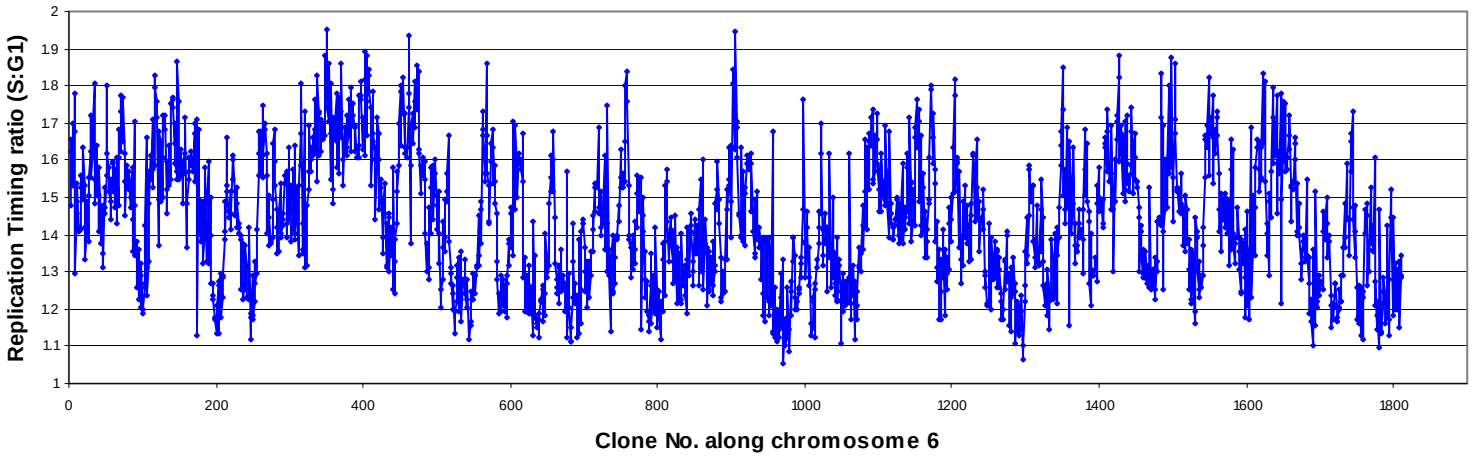
Figure 5.7: Replication Timing profile of chromosome 22 with other genome features. A: GC content, B: Exon density C: *Alu* repeat content D: LINE repeat content.

5.3.2: The Replication Timing of Chromosome 6.

Replication timing was assayed using a chromosome 6 tiling path array. In light of the reproducibility of measurements on the 1Mb and 22 tile path arrays, the number of replicates assayed for chromosome 6 was reduced to two. The average co-efficient of variation between the two replicates is 1.54%. The replication timing profile of chromosome 6 is shown in Figure 5.8. Replication timing is plotted against the order of the clone on the chromosome 6 tile path rather than absolute position because at the time of this work clones for some regions, for example in the MHC locus, had not been mapped onto the finished chromosome 6 sequence.

Linear regression was then performed on this data to assess the correlation between replication timing and sequence features. The relationship between replication timing and sequence features of the genome are shown in Figure 5.9. The linear regression statistics are shown in Table 5.5.

A


Replication Timing ratio (S:G1)
2
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1
0
200
400
600
800
1000
1200
1400
1600
1800
Clone No. along chromosome 6

B

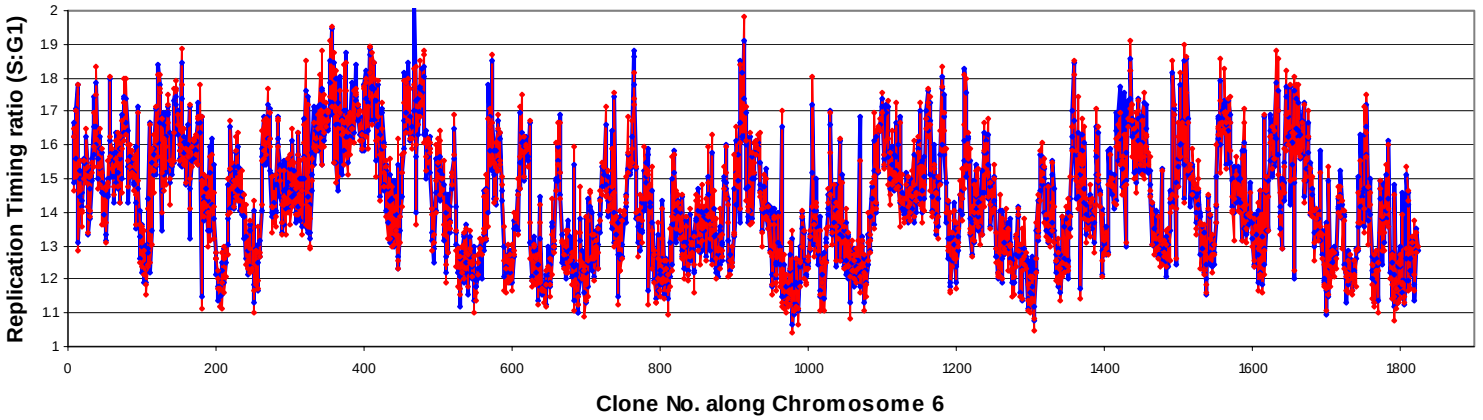

Replication Timing ratio (S:G1)
2
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1
0
200
400
600
800
1000
1200
1400
1600
1800
Clone No. along Chromosome 6

Figure 5.8: (Previous page) The replication timing profile of chromosome 6. A: Average replication timing profile of two array experiments. B: Replication timing profile of chromosome 6 showing reproducibility of the data. Blue: replicate 1, Red: replicate 2.

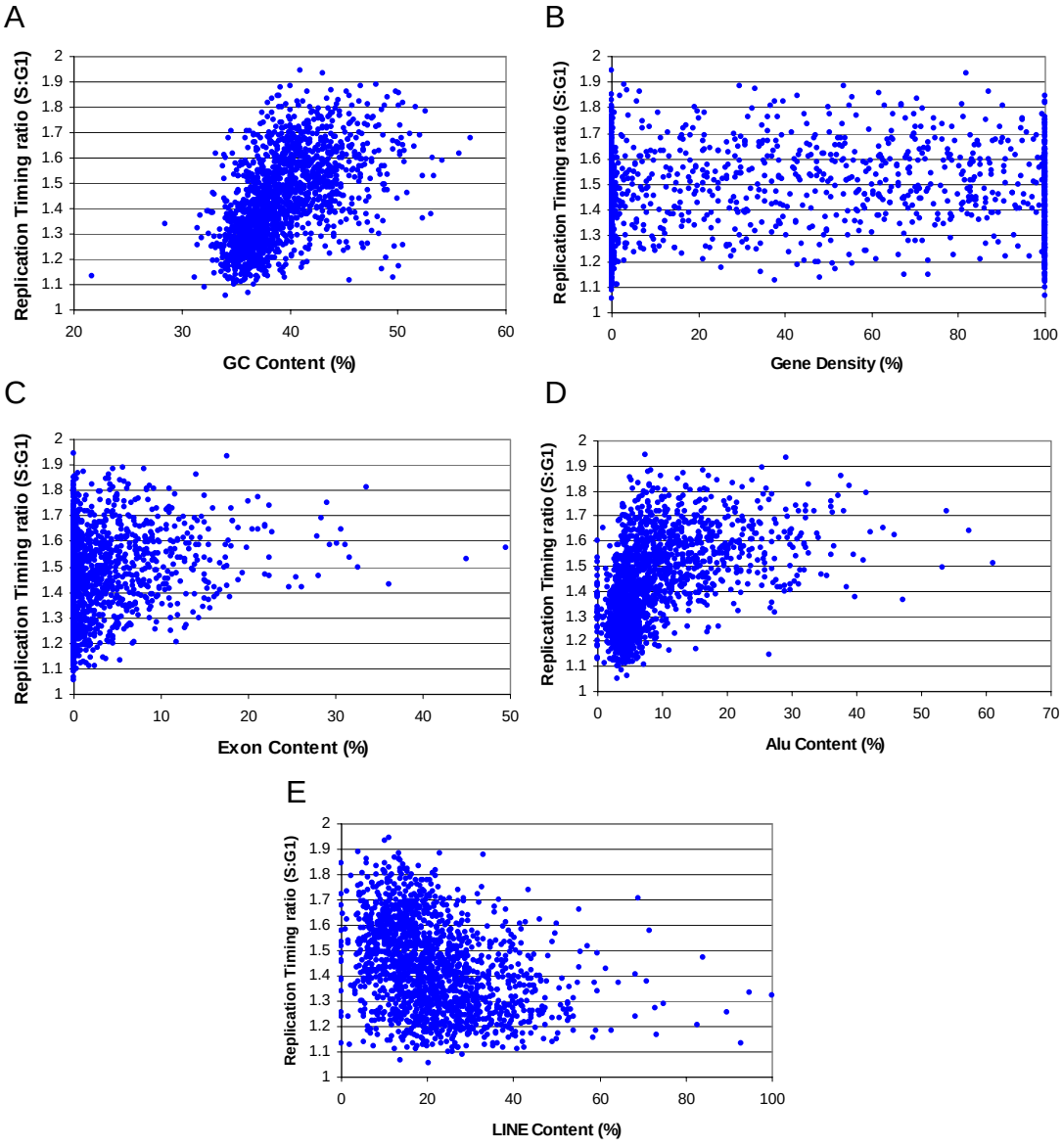


Figure 5.9: Correlation between replication timing and other genome features on the chromosome 6 tile path array. A: Correlation with GC content. B: Correlation with gene density. C: Correlation with exon density. D: Correlation with *Alu* repeat content. E: Correlation with LINE repeat content.

Table 5.5: Linear regression performed between replication timing and genome statistics at a 94Kb resolution on the chromosome 6 tile path array.

Genome Feature	**Regression Coefficient (x)**	**Intercept (y)**	**Correlation Coefficient (r)**
GC Content	0.024	0.48	0.54
Gene Density	0.001	1.40	0.22
Exon density	0.010	1.41	0.30
***Alu* Repeat Content**	0.010	1.34	0.48
LINE Repeat Content	-0.005	1.53	0.34

Unlike the chromosome 22 tile path data, but in common with the correlations seen at a 1Mb resolution and on a chromosome wide analysis, the best correlation was observed with GC content. The worst correlation observed was with gene density, with a correlation co-efficient of just 0.22.

5.3.3: The Replication Timing of Chromosome 1

Duplicate chromosome 1 tiling path arrays were used to assay the replication timing of this chromosome. The average co-efficient of variation between the two replicates was 2.24%. The replication timing profile of chromosome 1 is shown in Figure 5.10.

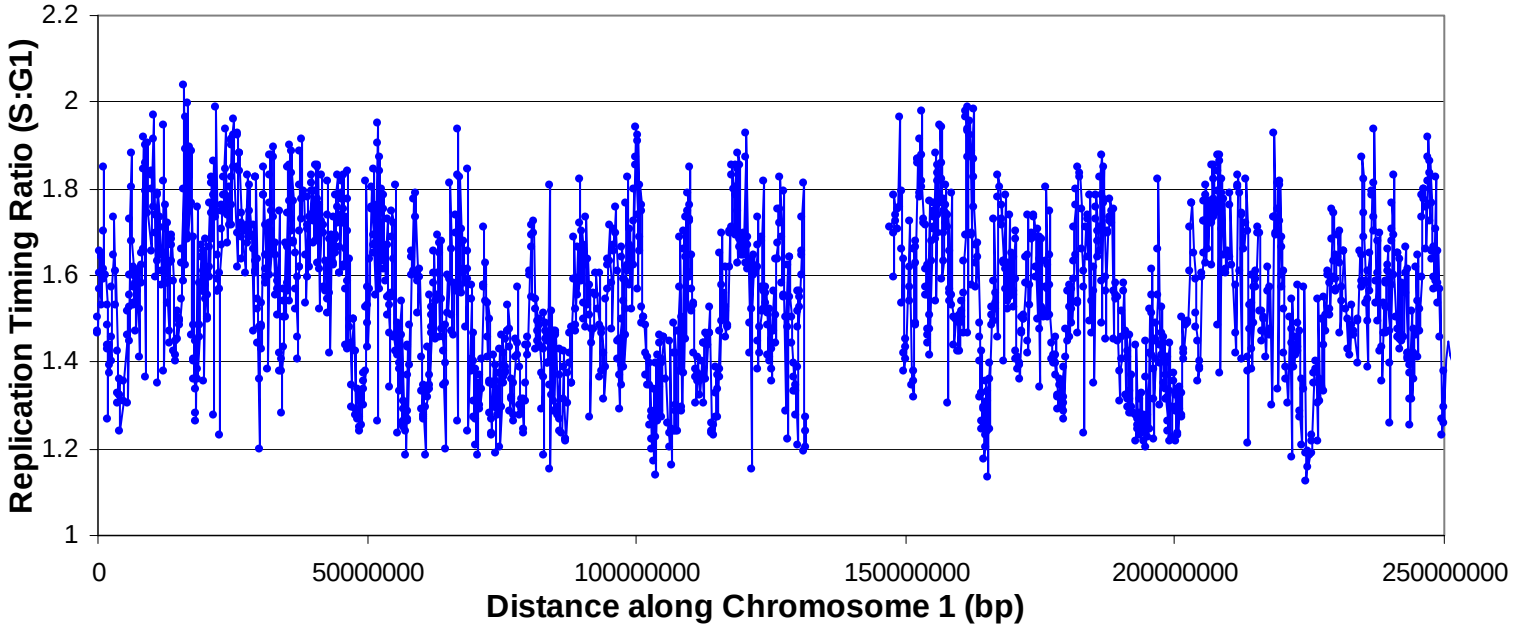


Figure 5.10: The replication timing profile of chromosome 1. The ratios reported are the average of two arrays.

Linear regression was then performed on this data to assess the correlation between replication timing and sequence features. The relationship between replication timing and sequence features of the genome are shown in Figure 5.11. The linear regression statistics are shown in Table 5.6.

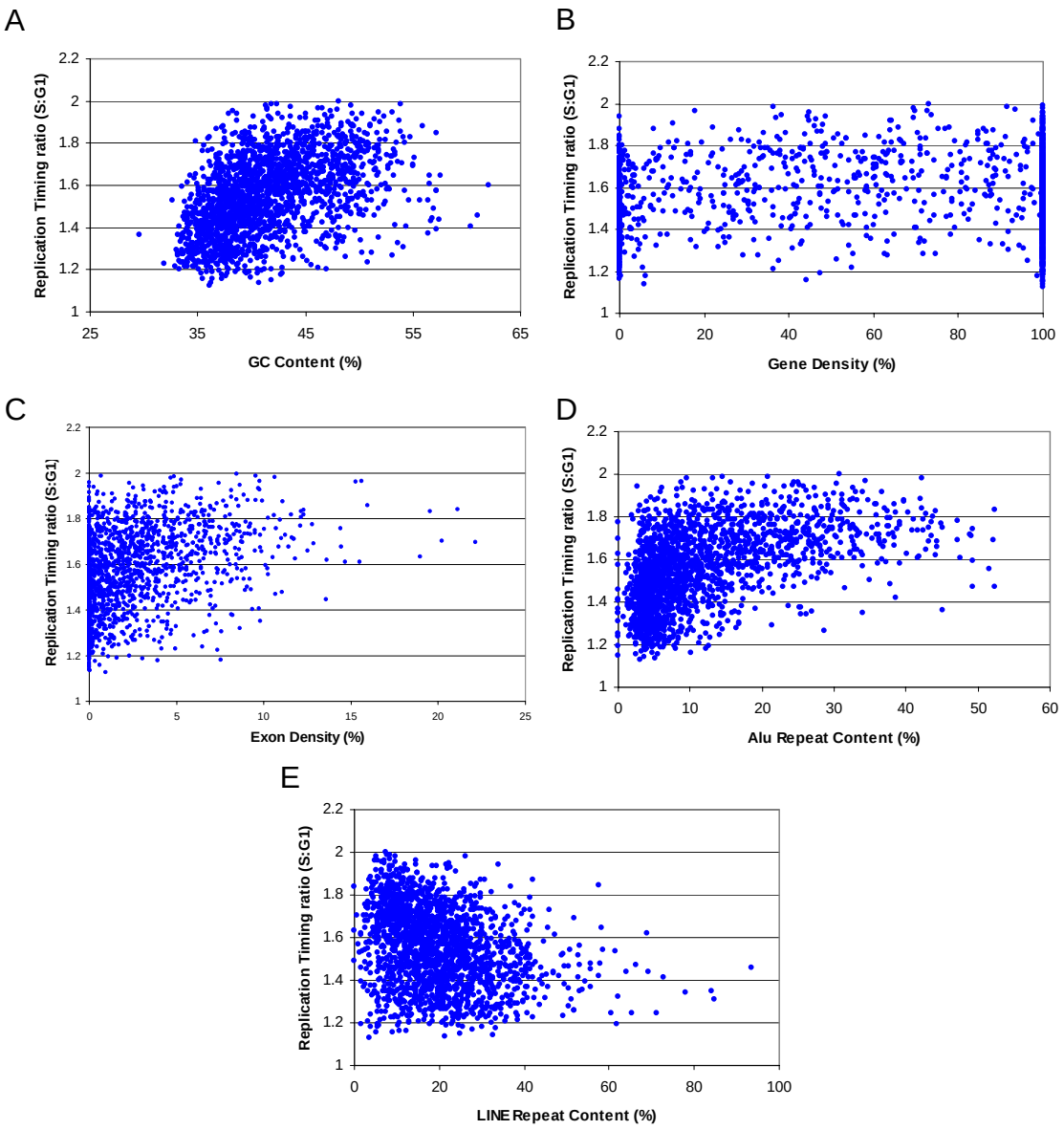


Figure 5.11: Correlation between replication timing and other genome features on the chromosome 1 tile path array. A: Correlation with GC content. B: Correlation with gene density. C: Correlation with exon density. D: Correlation with *Alu* repeat content. E: Correlation with LINE repeat content.

Table 5.6: Linear regression performed between replication timing and genome statistics at a 94Kb resolution on the chromosome 1 tile path array.

Genome Feature	Regression Coefficient (x)	Intercept (y)	Correlation Coefficient (r)
GC Content	0.016	0.88	0.45
Gene Density	0.0004	1.52	0.08
Exon density	0.024	1.50	0.40
***Alu* Repeat Content**	0.009	1.44	0.51
LINE Repeat Content	-0.005	1.64	0.30

As reported by the chromosome 22 array, the best correlation was seen with *Alu* repeat content, and again a poor correlation with gene density was found.

5.3.4 Comparison of Replication timing between two different lymphoblastoid cell lines.

The replication timing of two different lymphoblastoid cell lines was examined. S and G1 phase DNA was flow sorted from a HRC 575 (male) cell line and a HRC 160 (female) cell line. The S and G1 phase DNA from each sort was differentially labelled. The DNA from HRC 575 and HRC 160 was then hybridised to individual chromosome 22 tile path arrays.

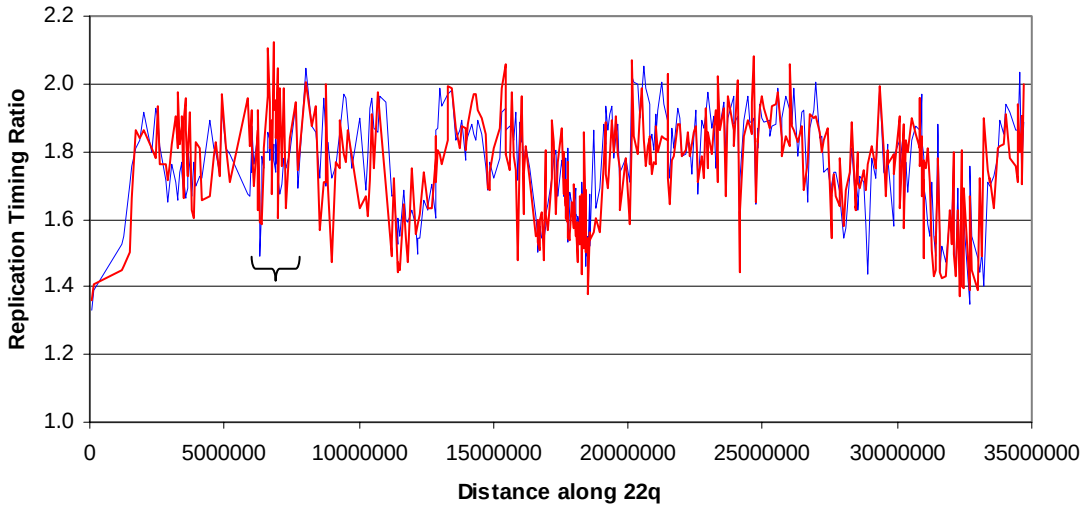


Figure 5.12: Replication timing profiles of two different lymphoblastoid cell lines. Blue: Male lymphoblastoid cell line (HRC 575). This cell line has a deletion in the

immunoglobulin light chain λ region, detailed in 4.2 and marked by the black scroll. Red: Female lymphoblastoid (HRC 160) cell line, with no deletion.

Few regions of replication timing difference can be seen outside the experimental variation of 0.1. The exception is the immunoglobulin light chain λ region, which has been shown to have a different copy number in each cell line, and a region 3Mb from the start of the q arm sequence. Regression analysis was performed on this data to compare the reproducibility of replication timing in different cell lines of lymphoid origin. The correlation coefficient was found to be 0.73 if regression was performed on the whole of 22q, and 0.79 if performed on the last 25Mb used for the sequence feature correlations.

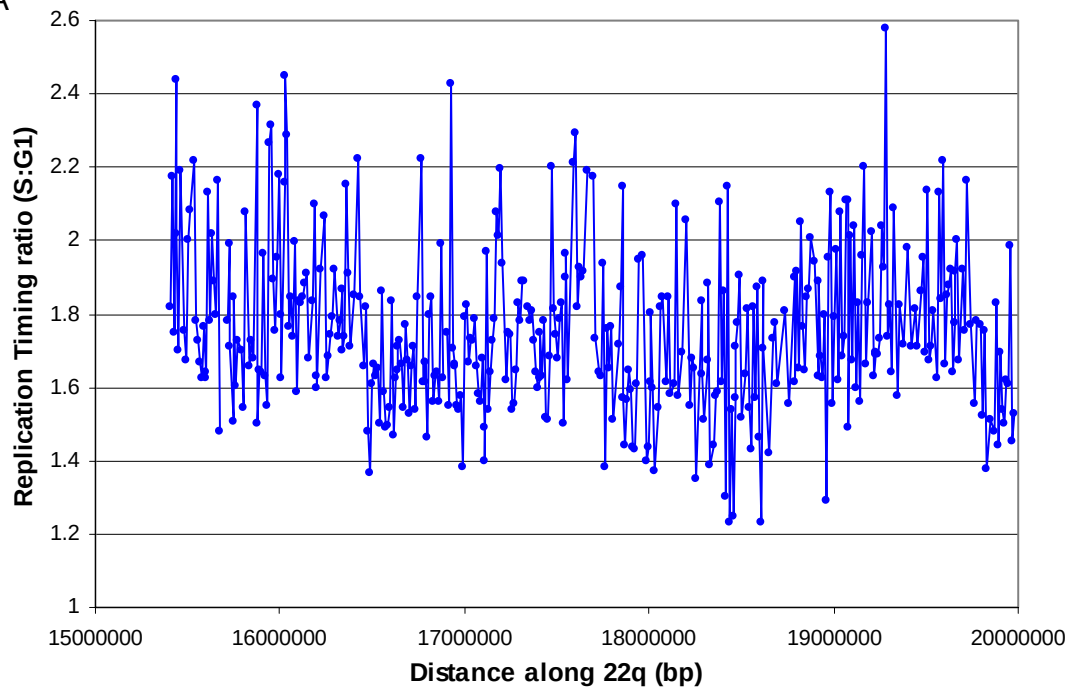
5.4: Assessment of Replication Timing at High Resolution Using an Array constructed with 500bp PCR Products.

An S:G1 hybridisation was performed on an array consisting of high resolution PCR products. In common with the self:self hybridisations described in section 4.6. the noise observed on the PCR product arrays showed a greater amplitude compared to the clone product arrays. The S:G1 hybridisations were repeated six times on separate arrays. The average coefficient of variation for each locus on the array was 14.47%. This is very similar to the coefficient of variation of 14.65% produced by the self:self hybridisation by the chromosome 22 products on the high resolution array reported in section 4.6.

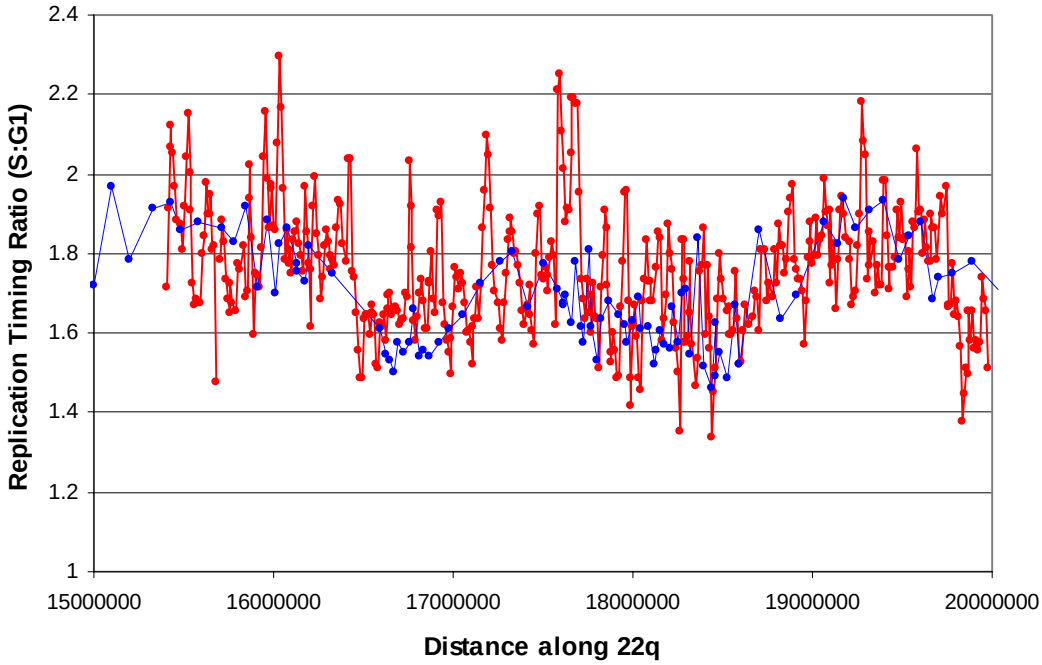
5.4.1: PCR Product array at 10Kb resolution.

The region of chromosome 22 15.4 - 20Mb along the q arm was assayed at a 10Kb resolution. This region was chosen because it included a transition from late-early replication when the timing was assayed on the 22q tile path array. The transition was identified as being approx 16.4Mb along 22q and was between clones dJ90G24 and cN38H9. The array was normalised to the average replication timing observed for the 15.5-20Mb region on the 22q tile path array (i.e 1.668). The replication timing profile for this region is shown in Figure 5.13.

A


Replication Timing ratio (S:G1)
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
15000000
16000000
17000000
18000000
19000000
20000000
Distance along 22q (bp)

B


Replication Timing Ratio (S:G1)
2.4
2.2
2
1.8
1.6
1.4
1.2
1
15000000
16000000
17000000
18000000
19000000
20000000
Distance along 22q

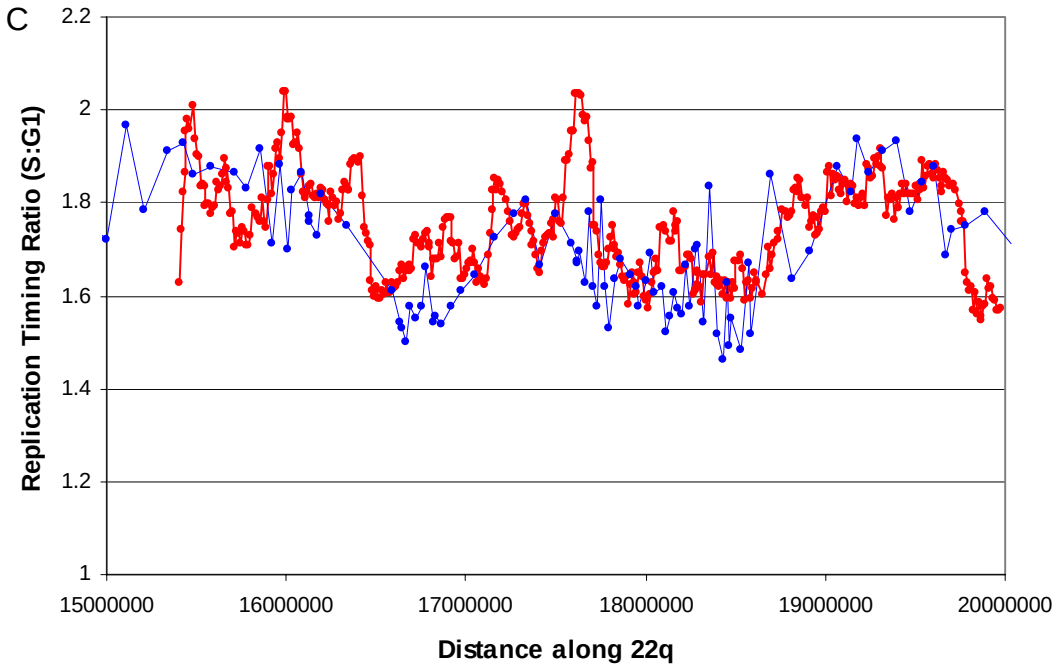


Figure 5.13: Replication timing profile of the region 15.4-20Kb along 22q at a 10Kb resolution. A: Replication Timing sampling all loci. B: Average replication timing utilising a 30Kb moving window. Red: Data from 500bp PCR product arrays. Blue:Data from the 22q clone array. C: Average replication timing utilising a 100Kb moving window. Red: Data from 500bp PCR product arrays. Blue: Data from the 22q clone array.

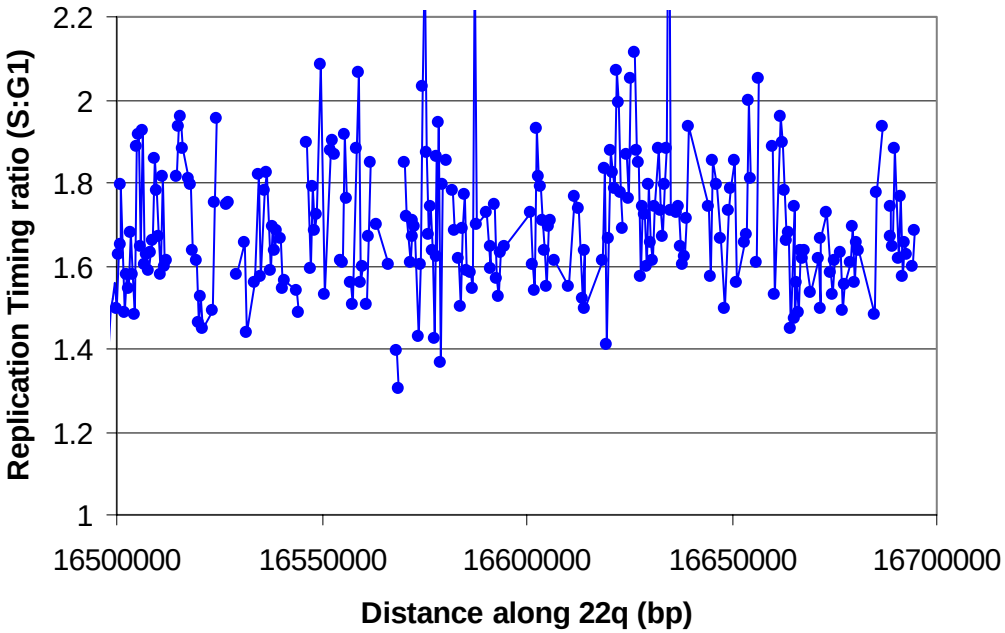
A transition in replication timing, from an early region to a late region can be observed 16.43-16.48 Mb along chromosome 22q. This shows a parallel with what is observed on the 22q tile path array, but narrows down the transition region from the 16.33-16.59Mb along chromosome 22q seen on the lower resolution tile path array. However transitions not observed on the 22q tile path array can be seen on this higher resolution array. A comparison between the profiles observed on the tile path and high resolution array can be seen on Figure 5.13.

The replication timing profile on the high resolution arrays shows a general correlation with those reported by the 22 tile path array. However there are also some inconsistencies, for instance, the early replicating region observed on the high resolution array between 17.8-17.9Mb is not detected by the 22q tile path clones within this region.

Chromosome X PCR products were also spotted onto the array for copy number change verification. The average replication timing of chromosome X is significantly later replicating than chromosome 22 (mean S:G1 ratio 1.38 as opposed to 1.75). The PCR products derived from chromosome X should therefore show a later replication than those from chromosome 22. This was found, with the average replication timing of the chromosome X products being 0.76, compared to 1.67 on chromosome 22. The reporting of a ratio less than 1, the high co-efficient of variation and the reporting of ratios above 2:1 on the chromosome 22 products, reflects that there remain problems with the detection of replication timing on these PCR product arrays. This is discussed further in section 5.7.7.

5.4.2: PCR product array utilising overlapping 500bp products.

The region 16.5-16.7Mb along chromosome 22q was sampled using overlapping 500bp products. The array was normalised as described in 4.2. The replication timing profile for this region is shown in Figure 5.14.



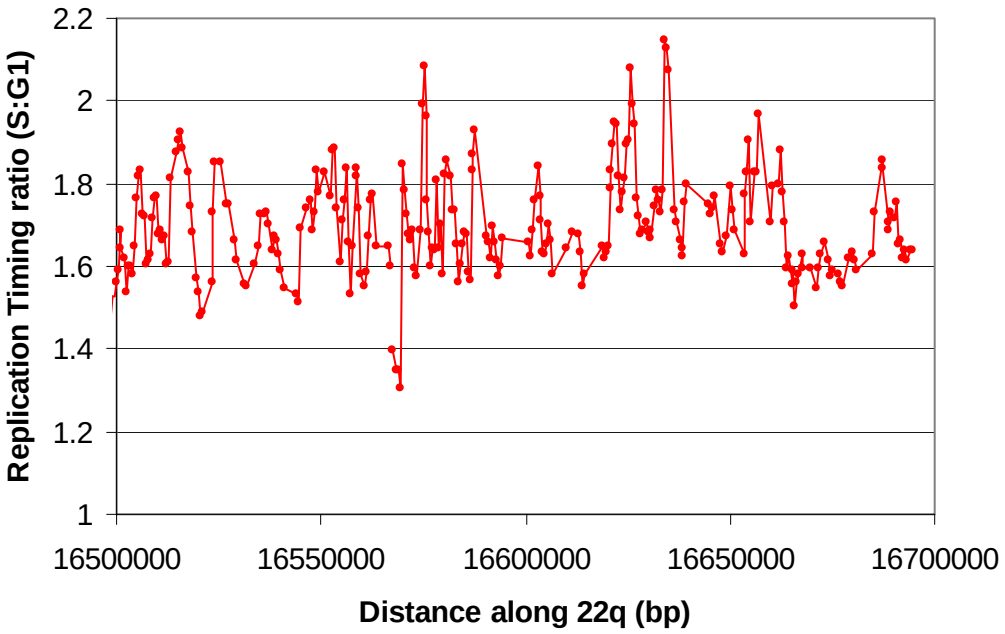


Figure 5.14: Replication Timing Profile of a region of 22q represented on the array by overlapping 500bp PCR products. A: Replication Timing sampling all loci. B: Average replication timing utilising a 1500bp moving window.

The profiles plotted in Figure 5.14 show gaps where the replication timing ratio is not reported. This is because sequences in these regions are not represented on the array

as insufficient unique sequence was available to allow the design of specific PCR primers.

A comparison between the profiles observed on the tile path, 10Kb resolution and a 500bp array can be seen in Figure 5.15.

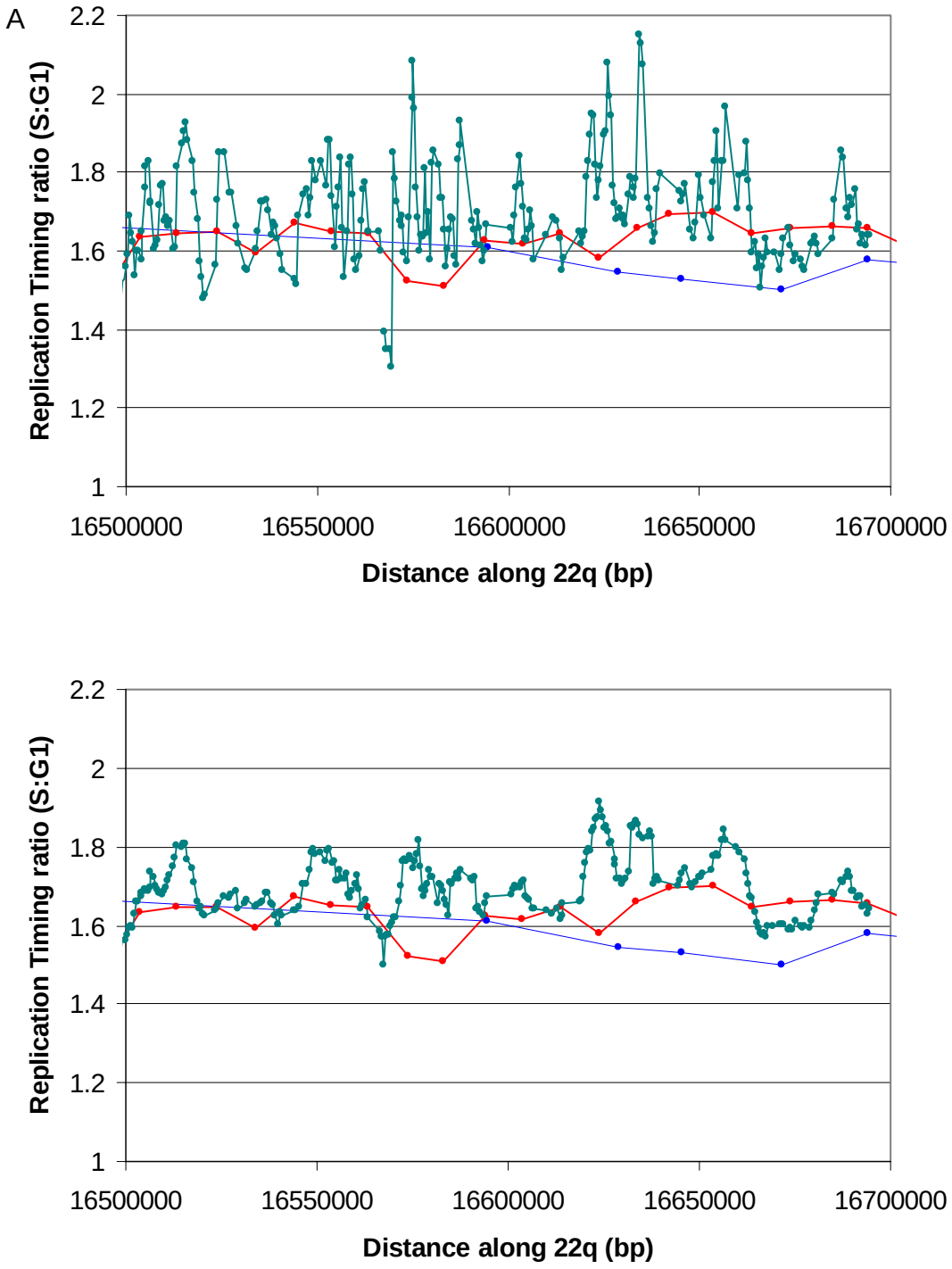


Figure 5.15: (Previous page) A: Comparison of the replication timing profile obtained on the 22 tile path array (blue) and a 10Kb resolution PCR product array (red) and a 500bp resolution array (green). B: Comparison of the replication timing profile obtained on the 22 tile path array (blue) and a 10Kb resolution PCR product array (red) and moving 5000bp window from the 500bp array (green).

The 500bp products show a profile that is different from that obtained with the 10Kb resolution array. However the 10Kb resolution array shows a good correlation with high resolution arrays at the points where 500bp products coincide with regions represented on the 500bp resolution array.

5.5: Correlation of assessment of Replication Timing by arrays with Replication Timing assessed by Quantitative PCR.

In order to validate my replication timing approach, I was able to compare our replication timing assay results from the 1Mb resolution array with a previously published independent analysis of chromosome arm 11q (Watanabe, Fujiyama et al. 2002).

Further corroboration of the method was performed by selecting clones represented on the 22q tile path array and calculating the difference in copy number between sorted S and G1 phase fractions by real time PCR.

5.5.1 Correlation with published quantitative PCR data on Chromosome 11q.

Wantanabe *et al* (Watanabe, Fujiyama et al. 2002) separated nuclei from a monocytic leukaemia cell line (46, XY) by flow sorting into 4 S phase fractions, extracted nascent DNA and then used semi-quantitative PCR to identify fractions enriched for specific STSs across the chromosome arm. DNA replicated in the 4th S phase fraction in the Wantanabe *et al* data would correspond to a late replication timing ratio on the array of 1.25:1 or below. Like-wise, replication in the 3rd S phase fraction would correspond to a replication timing ratio of between 1.25 and 1.5, *et cetera*.

The STSs on 11q were sequenced and remapped according to Build 31 of the human genome on the University of California, Santa Cruz website. The average spacing of each STS used in the Watanabe *et al* data is 300kb (Watanabe, Fujiyama et al. 2002), which is a resolution higher than that obtained with the 1Mb array.

The replication profile for both methods is shown plotted against chromosome 11 position in Figure 5.16. The replication profile of the two methods is very similar. Slight discrepancies can be seen at approximately 67, 95 and 111Mb along chromosome 11, where the replication timing reported by the arrays is later than that reported by the quantitative PCR method. A possible explanation for this discrepancy is that the two studies use a different cell line. The array data is obtained from a lymphoblastoid cell line with a normal karyotype and the Q-PCR data is obtained from a monocytic leukaemia cell line.

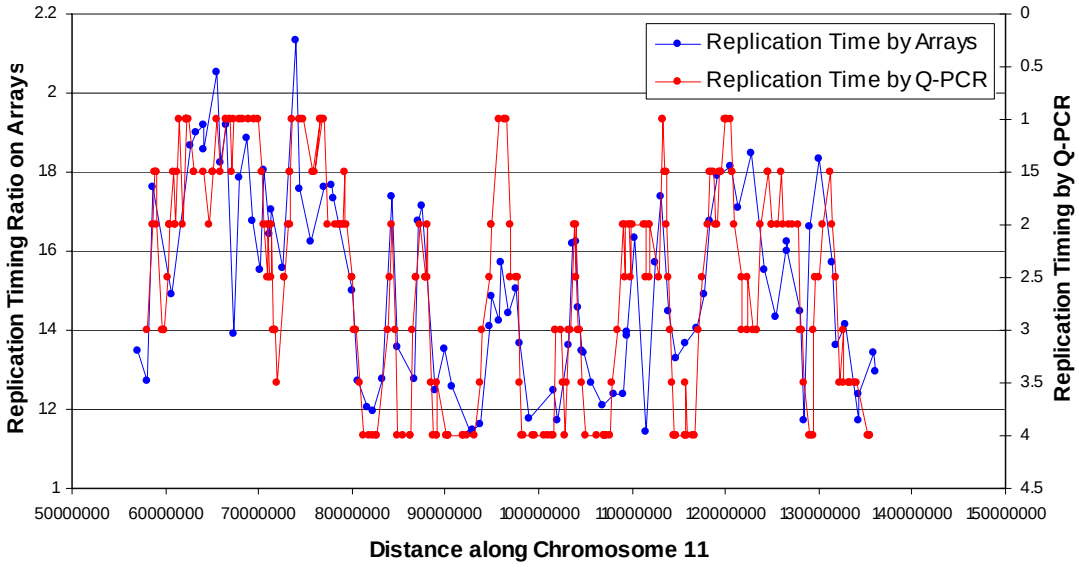


Figure 5.16: Replication Timing on 11q. Blue: Replication Timing reported by the 1 Mb array. Red: Replication Time reported by Wantanabe *et al.*

Data points within 100Kb of each other, from the two different methods were correlated as shown in Figure 5.17.

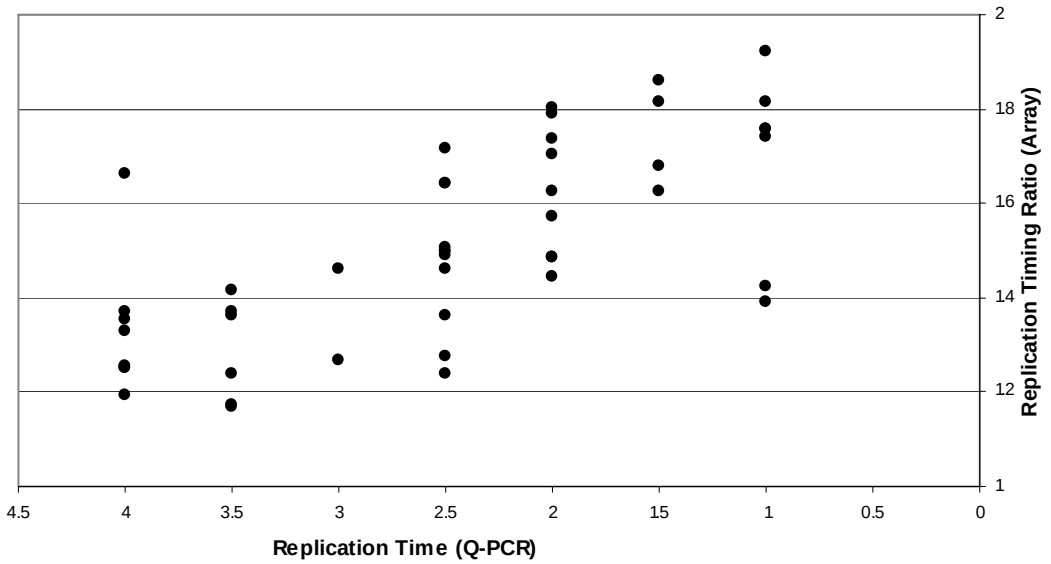


Figure 5.17: Correlation between quantitative PCR data and array data of loci within 100Kb of each other (y = -0.148x + 1.88 by linear regression).

A moderate-strong correlation (r=0.69) was found between the two methods. It should be noted that this is despite the use of two different cell types, albeit both lymphoid in origin.

The MHC region is located on chromosome 6 so replication timing data generated with the chromosome tiling path array can be used for comparison with previously published studies at this locus.

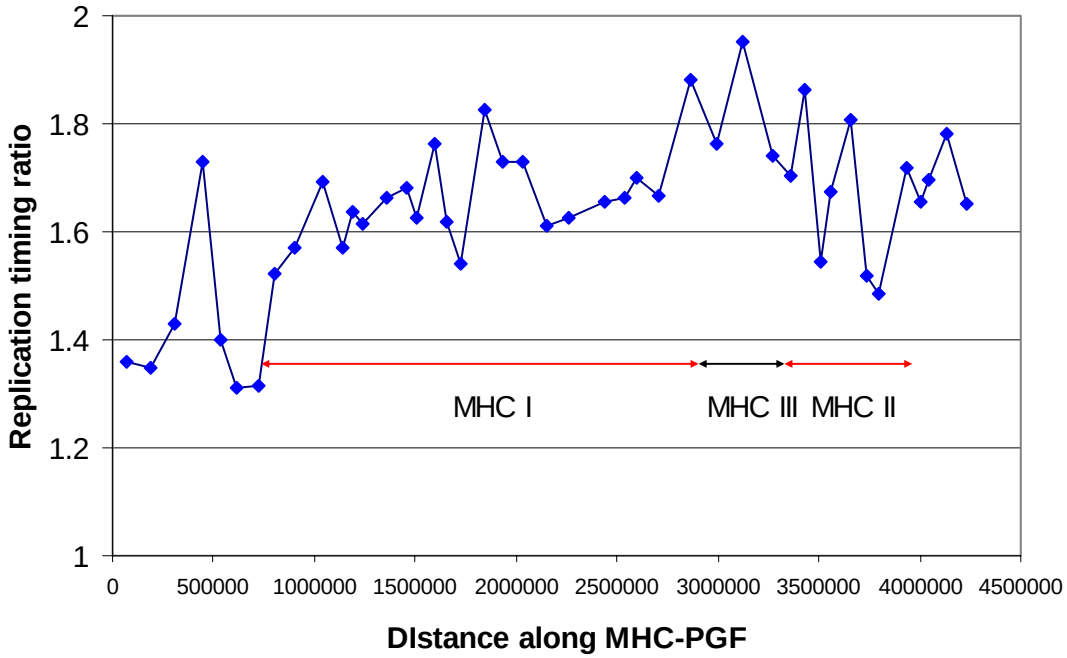


Figure 5.18: Replication timing data of the MHC region collected from the chromosome 6 tile path array

The replication timing data obtained by the arrays show a transition from early – late replication between MHC III and MHC II. This confirms what has been observed when this region was studied at high resolution by Tenzen *et al* (Tenzen, Yamagata et al. 1997) using a PCR based method.

5.5.2: Verification of replication timing by arrays by analysis by Quantitative PCR

To further verify our approach, four clones from chromosome 22 were chosen for analysis by real time PCR. These were one late replicating clone (cN69F4, position on X axis=1.38), two mid replicating clones (cE140F8 X=1.71 & cB13C9 X=1.64) and an early replicating clone (bK57G9 X=1.97). Primer pairs were designed every 10Kb along the clone. Each primer pair was assayed by real time PCR in quadruplicate. The average coefficient of variation was 7.2%. The S:G1 ratio for each primer pair was calculated and compared to the ratio obtained for the entire clone by array analysis (detailed in section 5.3.1) as shown in Figure 5.19.

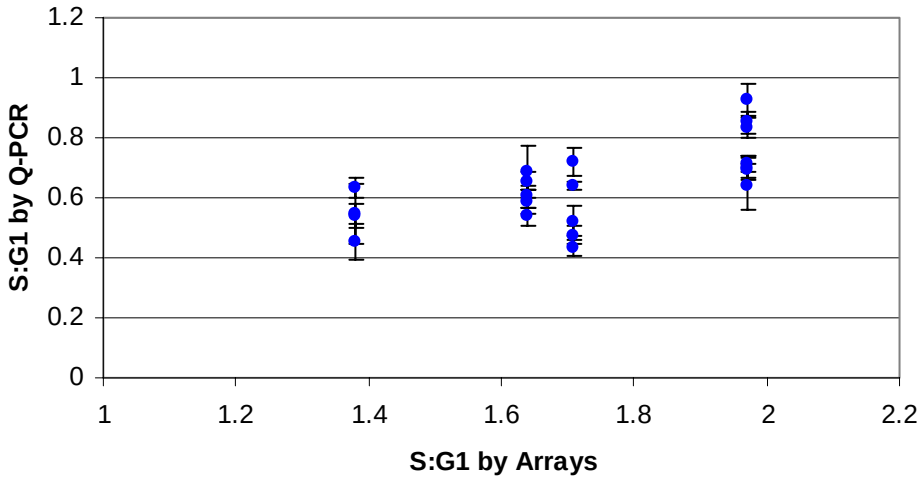


Figure 5.19: Comparison of S:G1 as determined by the replication timing arrays (X axis) and by quantitative PCR (Y axis). Y error bars show the standard error of each quadruplicate for the quantitative PCR experiment.

The PCR data was averaged over a clone (to make them more comparable with the array data) and the correlation co-efficient was calculated as 0.87. This supports the data presented in section 5.5.1 and reveals that using microarrays to assess replication timing produces comparable data to that produced by a method utilising real time PCR.

5.6: Replication time in flow sorted S phase fractions.

S phase was sorted into five different fractions as shown in Figure 2.2 and DNA was extracted. The four fractions were co-hybridised against G1 in four separate experiments. G1:G1 and G2/M:G1 hybridisations were also performed. These were normalised as described in Table 2.7.

The five earliest replicating, five latest replicating and five mid replicating clones were selected from the replication timing profile shown in Figure 5.5. The ratio obtained for each fraction is plotted in Figure 5.20.

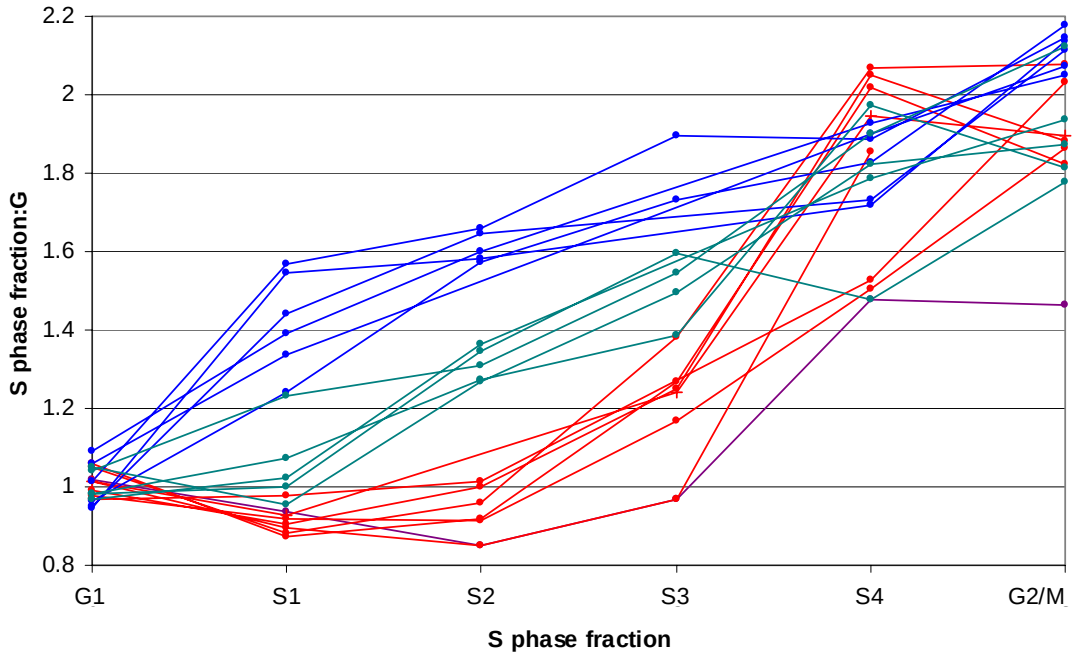


Figure 5.20: Ratio obtained for each S phase fraction when hybridised against G1. Four classes of clones are show. Blue: Early replicating, Green: Mid replicating, Red: Late replicating, Purple: centromeric, containing a large amount of common repeat elements. The clones used in this analysis are detailed in Table 5.7.

Table 5.7: Clones used for the analysis displayed in Figure 5.20

Class	Clones
Early replicating	dJ355C18, bK212A2, bK57G9, cN84E4, dJ1119A7, bK221G9
Mid replicating	dJ90G24, cE140F8, dJ15I23, bK243E7, cB13C9
Late replicating	cN2H8, cN29F4, cN22D1, cN69F4, bK262A13, bA191L9, cN129H9
Centromeric	cN14H11

Early replicating clones will double in copy number within the first fraction of the S phase sort (S1). They will then remain at a double copy number throughout the rest of S phase fractions. Conversely, late replicating clones will remain at a single copy number throughout early S phase fractions and double in copy number in the S4 fraction. Mid replicating clones will double in copy number in the S2 or S3 fractions. All clones should have replicated by the G2/M fraction

This is reflected in Figure 5.21. Early replicating clones increase in copy number in early S phase fractions, while late replicating clones increase in copy number in late S phase. Most clones show an average ratio of 2:1 when the G2/M fraction is ratioed

against G1. One clone in which this is not the case is the clone cN14H11 (purple on Figure 5.21). This clone is the most centromeric sequence clone of the q arm of chromosome 22 and is rich in common repeat elements. The incomplete suppression of these repeat elements may explain why the ratio reported for the G2/M:G1 hybridisation is only 1.46, instead of 2:1.

A region that showed a transition between an early and a late replicating DNA was also analysed. The region chosen was 26.8-28.0Mb along 22q. The results can be seen in Figure 5.21.

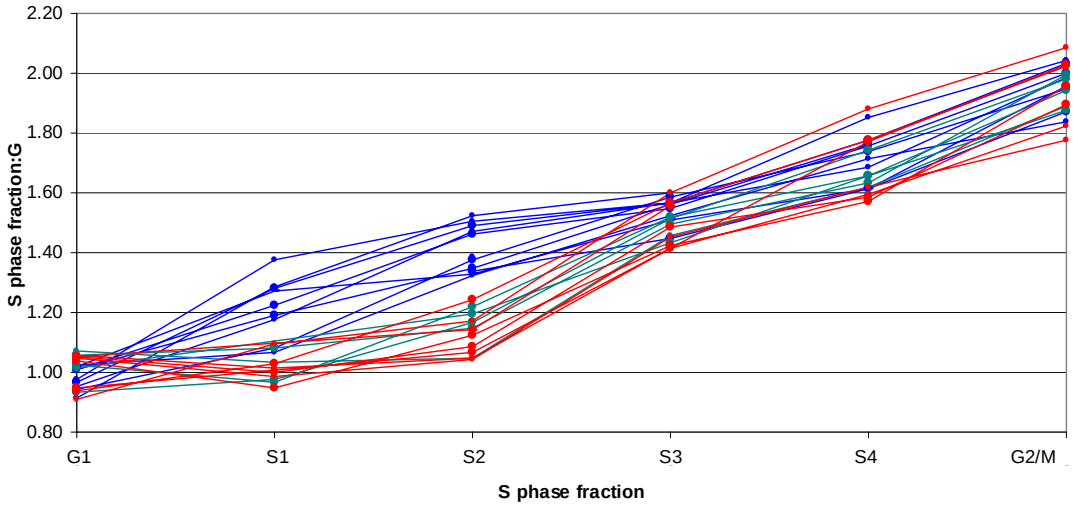


Figure 5.21: Ratio obtained for each S phase fraction when hybridised against G1. Blue: Early replicating side of transition, Green: Clones within the transition of replication timing, Red: Late replicating side of transition.

This analysis of consecutive overlapping clones confirms what is seen in Figure 5.22. Early replicating clones increase in copy number in early S phase fractions whilst the later replicating clones increase in copy number in later S phase fractions.

Replication timing analysis of the whole genome (section 5.2.1) shows that chromosome 22 is an early replicating chromosome, with an average replication timing ratio of 1.75:1. This is confirmed by a great majority of the chromosome 22 clones increasing in copy number in the S1 or S2 fraction of S phase. Conversely chromosome X is late replicating with a replication timing ratio of 1.38:1. Analysis of chromosome X clones on the array reveals that all clones representing chromosome X

increase copy number within fractions representing the latter half of S phase (Figure 5.22).

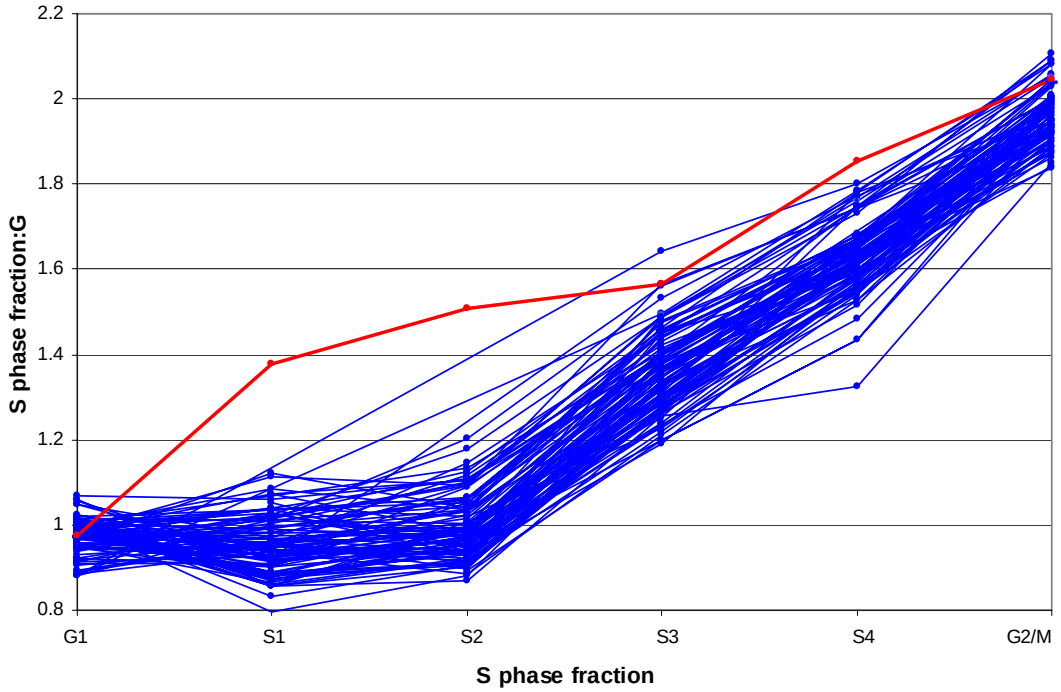


Figure 5.22: Ratio obtained on loci representing chromosome X for each S phase fraction when hybridised against G1 (Blue: chromosome X clones. Red: a typical chromosome 22 clone).

5.7: Discussion

The data presented in Sections 5.2-5.4 demonstrate how microarrays sampling genome sequence can be used to assess replication timing. Unlike conventional methods of assaying replication timing, genomic arrays report the replication timing of large genomic regions with a high accuracy. The spatial resolution of the method is only limited by the clone size and density of clones represented on the array. The work in this Chapter describes, for the first time, a high resolution replication timing analysis of a mammalian genome.

The replication timing was correlated with sequence features of the genome. Correlation coefficients for each feature are as reported in sections 5.2 and 5.3 and summarised in Table 5.8. Regions of early replication map to G light chromosomal bands, whilst regions of late replication map to G dark chromosomal bands.

Table 5.8: Regression co-efficients for correlations between replication timing and sequence features of the genome. (TP = tile path array). The strongest individual correlations are highlighted in red, whilst the weakest correlations are highlighted in blue. The best correlations are seen when all sequence features are combined, and are highlighted in green.

Genome Feature	**Chromosome Wide**	**1 Mb Chip**	**22 TP**	**6 TP**	**1 TP**
GC Content	0.96	0.7	0.22	0.54	0.49
Gene Density	0.89	0.35	0.19	0.22	0.08
Exon density	Not done	0.42	0.39	0.15	0.41
***Alu* Repeat Content**	0.9	0.56	0.45	0.48	0.51
LINE Repeat Content	0.72	0.4	0.34	0.34	0.3
Multiple Regression Analysis	Not done	0.76	0.57	Not done	Not done

5.7.1: Correlation between Replication Timing and Sequence Features.

Initial analysis of replication timing on a chromosome wide level shows that chromosomes of similar sizes exhibit very different replication timings. For example chromosomes 18 and 21 were very late replicating, whilst chromosomes 19 and 22 were early replicating. Chromosome 18 has already been shown to be late replicating and very gene poor. In contrast chromosome 19 is early replicating and gene dense (Zink, Bornfleth et al. 1999; Cross, Clark et al. 2000). Chromosome X which is abnormally rich in LINE repeats (IHGSC 2001), and the Y chromosome which includes a large amount of heterochromatin on the q arm, were both shown to be late replicating.

This initial large scale analysis also revealed that the gene deserts on chromosomes 13 (48-89Mb) and 14 (79-86Mb) defined by sequence analysis of the genome (IHGSC 2001) are late replicating. This is shown in Figure 5.23.

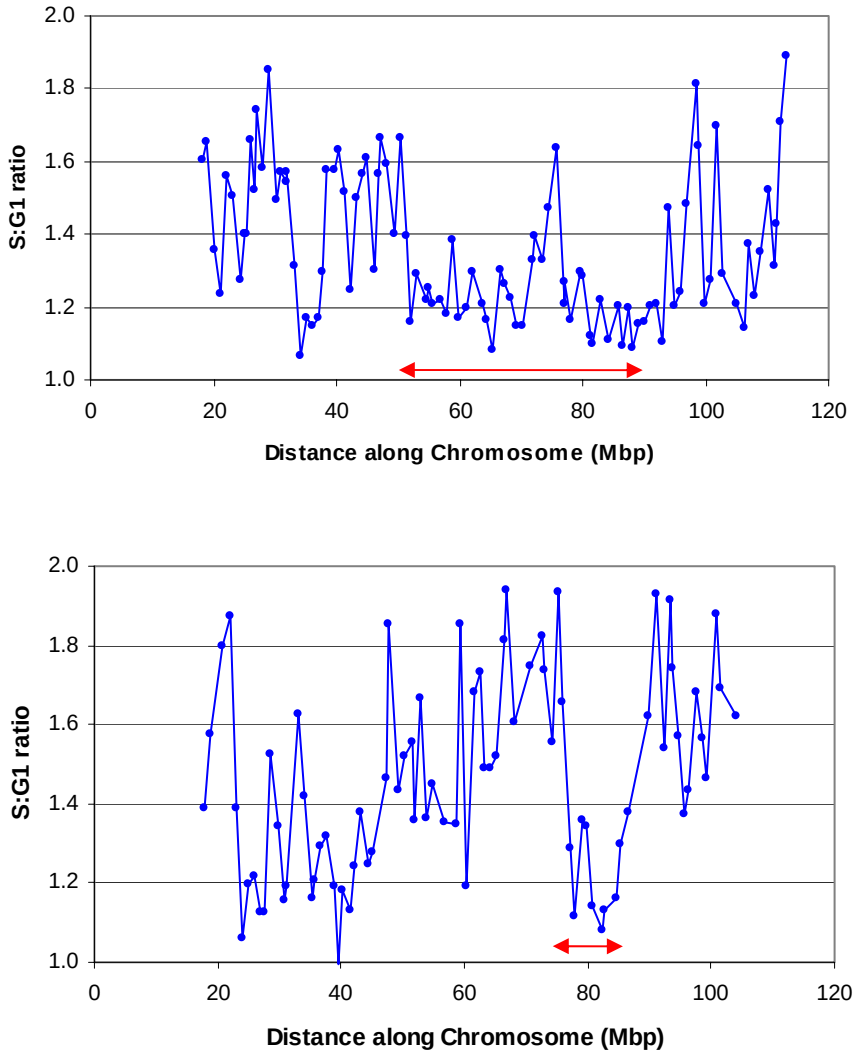


Figure 5.23: Replication timing of chromosomes 13 and 14. Gene deserts are marked with the red arrow.

Initial inspection of the regression coefficients shows that as the sampling resolution of the genome increases, the correlations with sequence features decrease. Therefore the chromosome wide correlations are much better than those using information from the tiling path arrays. As all the features are interrelated, as shown in Table 1.1, it is difficult to determine individually which features drive early replication. The effect of genome features on replication timing will be considered in turn below.

These results demonstrate that replication timing can be assayed at a tiling path level and that the replication timing can be correlated with other genome features. However the average resolution of the 22q tiling path array is only 78Kb and the average

replicon is thought to be approx. 40-100Kb (Nakamura 1986; Natale, Li et al. 2000) so to assay replication timing at the level of the replicon a higher resolution array needs to be used.

5.7.1.1: Correlation between Replication Timing and GC Content.

The best correlation on a chromosome wide basis was seen with GC content. This was also the case when the whole genome was sampled at a 1Mb resolution. The correlation between replication timing and GC content has been previously reported, including across an R/G band boundary (Strehl, LaSalle et al. 1997) and along entire chromosome arms (Watanabe, Fujiyama et al. 2002).

Analysis of the correlation with GC content on a tile path level gives disparate results. The correlation on chromosome 22 was weak, with a correlation coefficient of 0.22. The only feature on chromosome 22 showing a weaker correlation was with gene density. However on chromosome 6 GC content shows the strongest correlation with replication timing with a correlation coefficient of 0.54. Analysis of the replication timing and GC content of chromosome 22, in relation to chromosomal position (Figure 5.7), shows that, generally, the replication timing does follow the GC content of this chromosome. However, replication timing and GC content become uncorrelated at 22q13. This region is unusual in that while it is GC rich it is gene and *Alu* repeat poor. In this region it appears that gene density rather than GC content may drive replication timing.

The correlation with GC content on the chromosome 6 and 1 tile path arrays is quite good. It should be noted that chromosome 22 is very small. Small regions of difference between GC content and replication timing at 22q13 have a large effect on the correlation performed on all of 22q. The results in Table 5.8 show that GC content is important and may influence replication timing.

5.7.1.2: Correlation with Gene Density.

A weaker correlation was observed with measures of gene density. Two different measures of gene density were assayed; gene density (all intragenic DNA) and exon density (Exonic DNA only). Gene density was defined as the percentage of exonic and intronic DNA that is found within each clone. Exon density was defined as the percentage of exonic DNA (not that within introns) found within a clone. The correlation with gene density on a chromosome wide level was strong, with a correlation coefficient of 0.89 as reported in Table 5.8. However the correlation coefficient on the other arrays was poor. The correlation with gene density was the weakest seen on three of the four arrays analysed, with the correlation on chromosome 1 being just 0.08. On the chromosome 6 tile path array the correlation with gene density was the second weakest with a correlation coefficient of 0.22. In the case of the chromosome 6 tile path the weakest correlation was with exon density. In this case the correlation coefficient was 0.15.

The correlations with exon density are also weaker than those observed with other genome features. The correlation seen on the tile path arrays were variable, as with GC content. The correlation on the chromosome 1 tile path array was modest with a correlation coefficient of 0.41, however the correlation on the chromosome 6 tile path was very weak with a correlation coefficient of 0.15.

This analysis shows that that the correlation between replication timing and measures of gene density were weak in comparison to those with GC content.

5.7.1.3: Correlation with Common Repeat Elements.

Replication timing was also correlated with two types of common repeat element, *Alu* repeats and LINE repeats. In line with the other genome features the best correlations were seen at a chromosome wide level.

The correlations with *Alu* repeats were strong. The correlation between replication timing and *Alu* repeat content exhibited either the strongest or second-strongest correlation at all resolutions tested.

The correlation between replication timing and LINE repeats was the poorest of all the sequence features investigated at a genome wide level. The correlations at a 1Mb and tile path resolution were all somewhat similar with the coefficient correlations ranging from 0.3-0.4.

When looking at the correlation with LINE content on a chromosome wide level an outlier can be identified. (Figure 5.2D at 32.8, 1.38); this is the locus that corresponds to chromosome X. Chromosome X is known to be unusually rich in LINE repeat elements (IHGSC 2001). If this point is removed from the analysis the correlation co-efficient is 0.88, which is similar to the correlation co-efficient of the other features.

Analysis of the correlation between replication timing and repeat content shows a strong positive correlation, whilst a negative correlation was observed with LINE repeat content. This is consistent with the characteristics of active and inert chromatin as documented in Table 1.1.

5.7.1.4: Inter-correlation between replication timing and sequence features.

Statistical analysis performed in collaboration with Richard Mott (Wellcome Trust Centre for Human Genetics, Oxford) showed, by multiple regression analysis, that the genome features investigated were highly correlated with each other.

The multiple regression analysis showed that the correlation coefficient between replication timing and all the sequence features explored was 0.75 when the genome was sampled at a 1Mb resolution, a small but highly statistically significant ($P<10^{-16}$) improvement over the 0.70 correlation with GC content data alone. Multiple regression on chromosome 22 revealed a correlation with all sequence features of 0.57. This a considerable improvement over the correlation with the best single sequence feature (*Alu* – 0.45, $P<10^{-16}$).

The correlations reported, are all indicative of transcriptionally active, open forms of chromatin being important for early replication. A further investigation between

replication timing and euchromatic, transcriptionally active chromatin is reported in section 6.2 and discussed in section 6.5.1.

5.7.2: Correlation between Replication Timing and chromosomal bands.

5.7.2.1: Correlation with Giemsa banding.

It is widely acknowledged that R bands (GC rich) in mammalian chromosomes replicate in the first half of S phase and G bands (GC poor) replicate late (Ganner and Evans 1971; Dutrillaux, Couturier et al. 1976; Holmquist, Gray et al. 1982). The correlation of replication timing with high resolution G banding is shown in Figure 5.24. The replication timing profile at a 1Mb level of chromosome 6 shows by visual inspection that G dark regions generally replicate late, such as those 48-51Mb and 93-96Mb along the chromosome, and G light bands replicate early, such as those 27-46Mb and 105-113Mb along the chromosome. These correlations can also be seen at a tile path resolution on chromosome 22 with the G dark bands 16.6-18.8Mb and 112.2-12.5Mb along the chromosome replicating late, conversely G light bands located 33.6-34.7Mb and 12.9-15.9Mb along the chromosome replicate early. The associations with G banding cannot be exact due to the different condensation levels of G dark and G light regions of metaphase chromosomes.

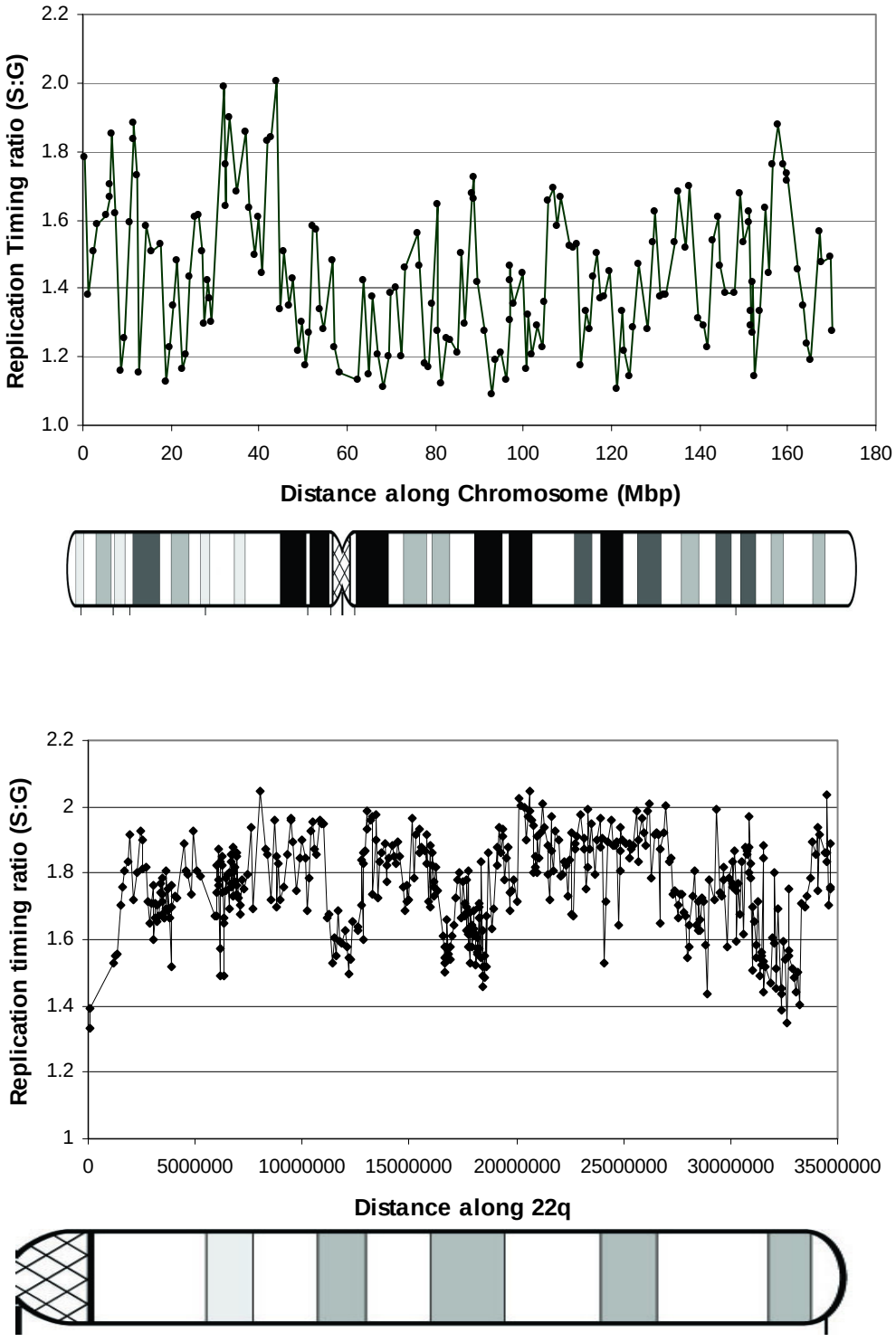


Figure 5.24: Comparison between replication timing ratio and high resolution giemsa banding of chromosomes (resolution = 850 bands, (Francke 1994)). (A) Chromosome 6 at a 1Mb resolution (B) Chromosome 22q at a tile path resolution

Statistical analysis on the replication timing ratios of clones that map to dark and light bands reveal there is a significant difference between the replication timing of sequence located in dark bands and sequence located in light bands.

An unpaired T test with Welch correction (to allow for the difference in means in the two populations when the variances are unequal) was performed on the chromosome 6 data obtained from the 1Mb array. Each locus on the array was assigned in either a dark or light band. The average replication timing of loci in light bands was 1.534 (standard deviation = 0.195) whilst the average replication timing of loci in dark bands was 1.334 (standard deviation = 0.164). The difference in replication timing observed was highly statistically significant with a P value less than 0.0001.

The same analysis was performed on the chromosome 22 tile path array data. The average replication timing of loci in light bands was 1.808 (standard deviation = 0.125) whilst the average replication timing of loci in dark bands was 1.677 (standard deviation = 0.153). The difference in replication timing observed was also highly statistically significant with a P value less than 0.0001.

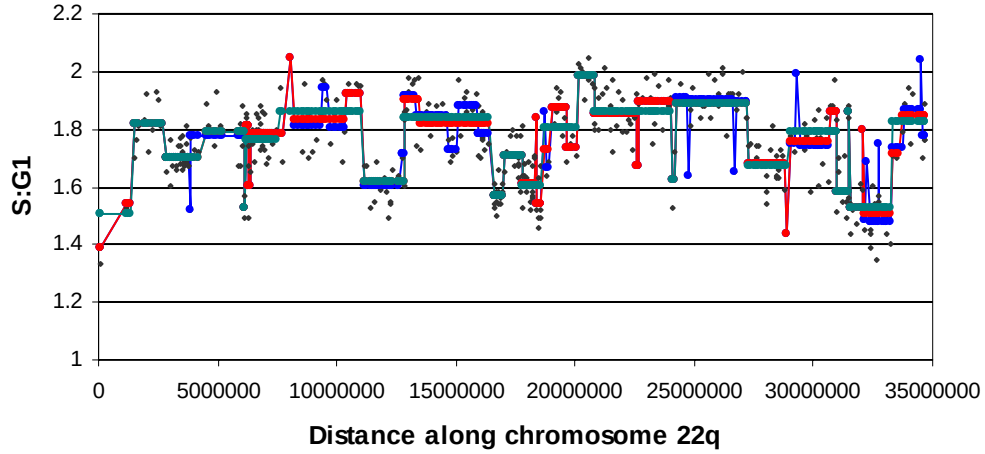
5.7.2.2: Regions of co-ordinated replication.

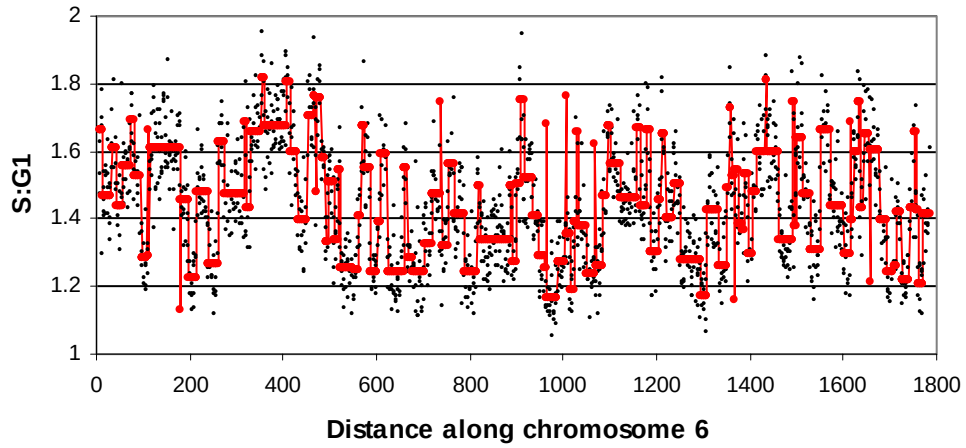
In order to assess the patterns of replication timing observed in the plot of the tile path data, we attempted to identify regions of similar replication timing and regions which differed significantly in replication timing from adjacent stretches in chromosomes 22, 6 and 1. A perl script was purpose written by Richard Mott at the Wellcome Trust Centre for Human Genetics for analysis of the replication timing data produced by the arrays. The program (detailed in Appendix 7) was used to find the optimal segmentation of the chromosome tile path data. Although the degree of segmentation observed can be adjusted by altering the segmentation penalty values, *B*, and it is not completely clear what a biologically meaningful value of this parameter should be; the analysis has the effect of delineating the patterns that are indicated by visual inspection.

This analysis was performed altering the segmentation penalty values (B), This was executed only on the 22q tile path data and indicated that the patterns of segmentation in the data was highly non-random, with $P < 0.001$.

Results of this analysis for a series of representative values of *B* are shown in Figure 5.25a for chromosome 22q. This illustrates that chromosome 22 has clear segments of consistently very early replicating DNA stretching over several megabases. Interspersed within these are megabase sized segments of later replicating DNA. Transitions between segments of early and late replicating areas of chromosome 22 (and vice-versa) are observed between data points whose midpoints are less than 160Kb apart (e.g. at ~11100000bp and ~12700000bp) suggesting disparate replication timing of adjacent replicons.

The statistical analysis described above was also performed on the data obtained from the chromosome 6 and chromosome 1 tile path arrays, using an intermediate segmentation penalty value. Consistent with the analysis of the chromosome 22 data, megabase sized segments of early, or late, replicating DNA could be identified. This is shown in Figure 5.25b and 5.25c.





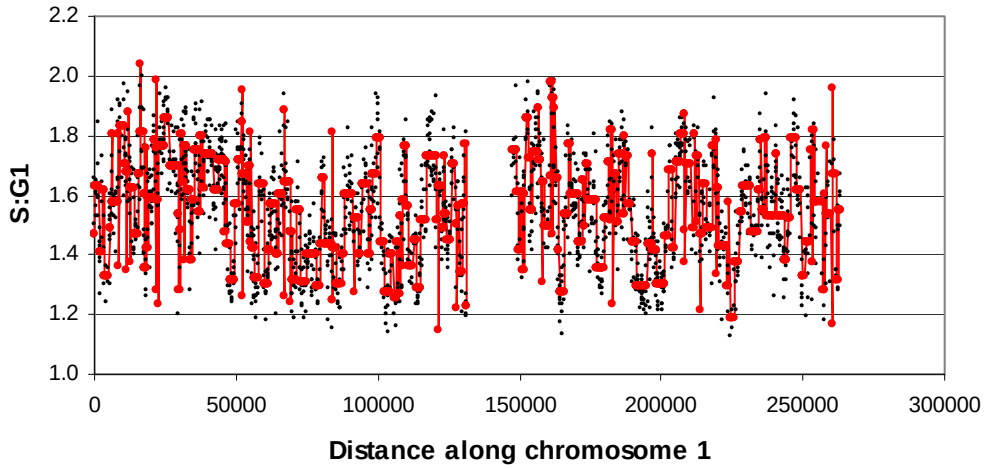


Figure 5.25: Statistical analysis to identify regions of the genome with similar replication timing. A: The graph shows the results of three runs of segmentation on the chromosome 22q data using representative segmentation penalty score (B) of 0.02 (blue), 0.04 (red) and 0.06 (green). Segmentation runs are plotted on top of the raw

replication timing data (black circles). B: Chromosome 6 (B = 0.04) C: Chromosome 1 (B = 0.04)

The comparison of the replication timing profiles with the banding patterns of chromosomes and the statistical analysis described above detailing the optimal segmentation of the replication timing data, showed that regions of the chromosome stretching over several megabases, replicate at similar times. The correlation with the giemsa banding of the chromosome suggests a link between replication timing and GC content. The identification of regions of several megabases that replicate at the same time suggest that groups of adjacent replicons replicate together. This is consistent with the observations made when chromosomes were studied by pulse labelling with BrdU (Dutrillaux, Couturier et al. 1976; Drouin, Lemieux et al. 1990; Cohen, Cobb et al. 1998), as described in section 1.3.

5.7.3: Rate of Replication

The replication timing data obtained from the 1Mb array can be used to assess the rate of genome replication. For this, S phase was divided into centiles based on S:G1 ratio. The number of loci replicating in each centile was counted and the cumulative number of loci replicated was plotted against the proportion of S phase completed (Figure. 5.26). Replication appears to start slowly, but increases to a linear rate of replication at about a third of the way through S phase finally again appearing to slow at the end of S phase. The slow initial rate of replication is supported by the shape of the distribution of S phase as measured on the flow cytometer (see S phase sorted fraction in Figure. 2.2) where there is a higher frequency of nuclei with lower DNA content. This implies that the DNA content of nuclei increases more slowly at the start of S phase and we can infer that either the frequency of the initiation of replication and/or the length of replicons are reduced during this period. The rate of replication then increases to a linear rate at about a third of the way through S phase. During this linear stage approximately 14% of the genome is replicated during each tenth of S phase. The replication rate slows towards the end of S phase. The slow rate of replication at the end of S phase cannot be explained from the cell cycle profile which displays a relatively even frequency of nuclei with increasing DNA content from the middle of S phase onwards. As most heterochromatin will replicate during this late

stage, and heterochromatic regions are not represented on the 1Mb genome wide arrays, the rate of replication for this final part of S phase is likely to be underestimated.

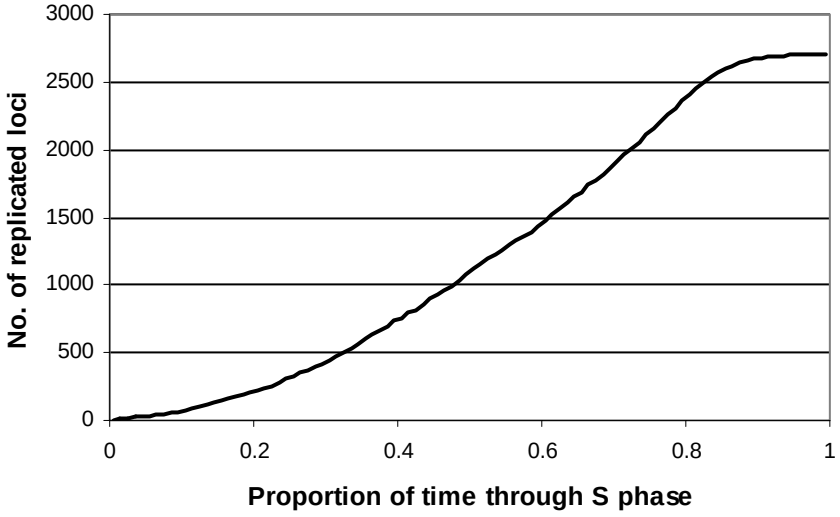


Figure 5.26: The rate of replication during the S phase of the cell cycle. Rate of replication is indicated by the slope of the curve plotted.

5.7.4: Comparison with other arrays assessing replication timing and limitations of the method.

Microarrays have previously been used to assess replication timing in the yeast (Raghuraman, Winzeler et al. 2001) and in *Drosophila melanogaster* (Schubeler, Scalzo et al. 2002). However the method used here is subtly different in a number of ways.

Firstly, the two previous studies have focused solely on coding regions of the genome. The study on yeast used the high density oligonucleotide array produced by Affymetrix. This chip represents each *Saccharomyces cerevisiae* open reading frame with up to 20 oligonucleotide sequences on the array. No regions outside the open reading frame were included (Raghuraman, Winzeler et al. 2001). The study on *Drosophila* used a cDNA array. The array was constructed to represent 5,543 expressed sequence tags from *D. melanogaster*. Both previous studies have therefore not assayed any non-coding regions of the genome, however, in yeast and *Drosophila* there is less non-transcribed DNA than is found in the human genome. The study

described in sections 5.2-5.3 assayed the human genome with DNA prepared from sequencing clones. This ensures both coding and non-coding regions of the genome are sampled. The representation of non-coding regions of the genome enable correlations with other sequence features such as GC content and repeat content to be performed. The inclusion of non-coding sequence also ensures there is an unbiased representation of sequences found in open and closed chromatin. Studies that only assay the replication timing for coding regions of the genome will not assay the replication timing of transcriptionally inert 'closed' chromatin and therefore any conclusions drawn are going to be biased towards what is found in transcriptionally active 'open' chromatin. To produce a complete understanding of replication timing both coding and non-coding regions of the genome should be sampled.

Secondly, the way replicating DNA is identified and extracted for application onto the array differs in each different assay. The methods used are described in 1.6.2. Briefly, for the yeast experiment, newly synthesised DNA was labelled with light carbon and nitrogen isotopes, in a background of DNA labelled with heavy isotopes. Post synchronisation samples were collected throughout S phase and a caesium chloride density gradient was used to separate newly replicated DNA from non replicated DNA. These were differentially labelled and co-hybridised to the array. The *Drosophila* experiment utilised cells pulse labelled with BrdU. The BrdU was incorporated into nascent DNA. The nuclei were then stained with propidium iodide and flow sorted by their PI intensity into an early S phase fraction and a late S phase fraction. The newly replicated DNA from each fraction was isolated by immunoprecipitation, amplified and differentially labelled using PCR.

Both these methods have the limitation that to obtain the replication timing ratio, one S phase fraction is ratioed against a different S phase fraction. This means that both methods may under-report very early replicating DNA sequences. A mean of the early S phase fraction is used as the early replicating reference point. DNA that replicates before this point will not be detected by the array. Also early replicating DNA may not be sufficiently labelled with BrdU to be detected. Using the method described in this thesis, the ratioing of S phase DNA against G1 phase DNA enables early replicating DNA sequences to be detected. This separation of S from G1 phase could not be achieved using the published *Drosophila* cell sort profile as the coefficient of

variation for the G1 and G2 peaks were too high. The S phase sort would therefore be contaminated from DNA from both the G1 and G2/M fractions of the cell cycle. This would affect the replication timing ratio reported. A comparison between the two flow sort profiles can be seen in Figure 5.27A

One limitation of the assessment of replication timing on arrays is the purity of the sort. Any contamination of S phase cells in the G1 fraction means that DNA from very early replicating loci may be present in the G1 fraction hybridised to the array. As a result, the true replication timing ratio of 2:1 may not be reported. Contamination of the S phase fraction with G1 phase cells will reduce all the S:G1 phase ratios. To verify the flow sorting purity, flow sorted fractions were passed back through the flow sorter and the degree of contamination was measured. This is shown in Figure 5.27B.

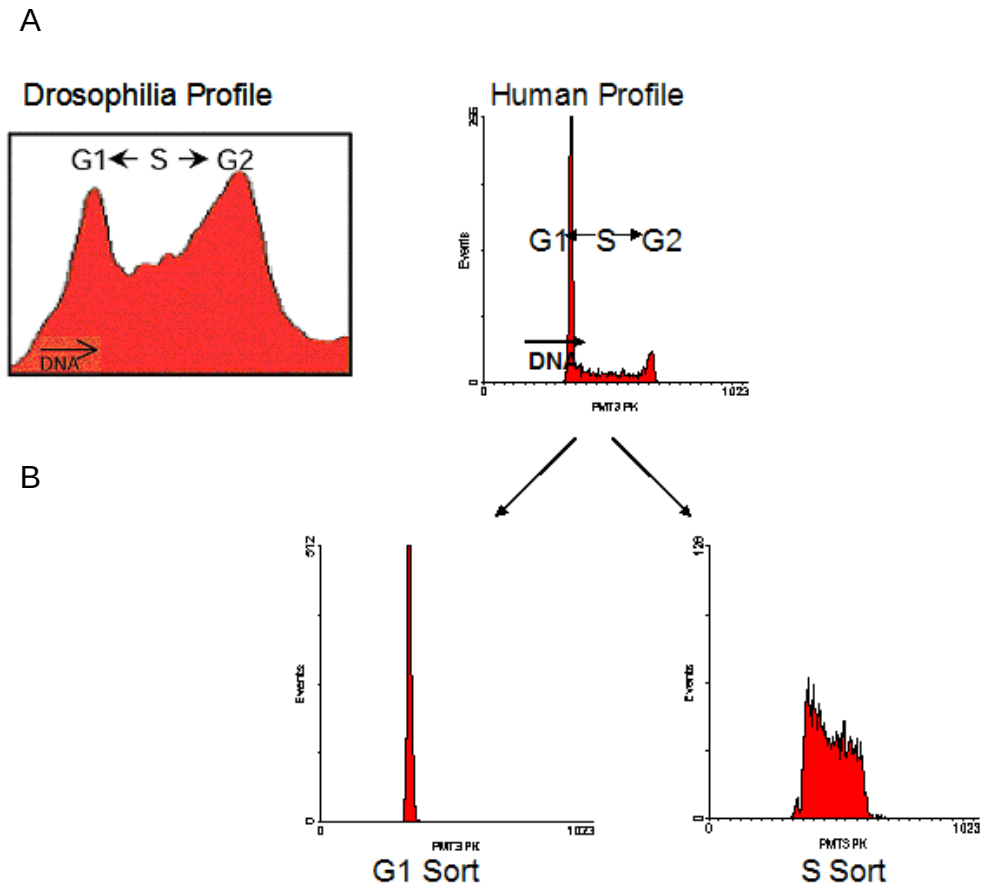


Figure 5.27: A: Comparison between *Drosophila* flow sort profile from (Schubeler, Scalzo et al. 2002) and the human lymphoblastoid flow sort profile obtained from sorting HRC575. B: Purity of the flow sort showing the G1 and S phase fractions.

Figure 5.27B shows that the flow sort is very pure, however there is a small amount of contamination and this will mask high S:G ratios of DNA which replicates very early in S phase and on the boundaries of G1 and S phase.

This contamination may also lead to inaccuracies in the normalisation of the array. The application of a curve fitting model to the cell cycle profile obtained in Figure 2.2 allowed extraction of best fit approximations of the distributions of G1 and S phase. The cell cycle analysis program Cylchred (Ormerod, Payne et al. 1987; Watson, Chambers et al. 1987) was used to estimate that within the S phase sorting window there are no contaminating G1 nuclei. However within the G1 phase sorting window there are approximately 2% of nuclei in very early S phase. In addition, approximately 4% of the earliest S phase nuclei are not represented within the S phase fraction. The consequence of this level of inaccuracy of sorting is that for a few very early replicating loci the theoretical replication timing maxima of 2.0 would be reduced to 1.93.

The purity of the flow sort is integral to the accuracy of the method and has proved a limitation when other tissues have been assayed. The evaluation of replication timing in other cell lines involves the culture of adherent cells. In these cell types I found that the Hoechst staining, vital for the sorting into G1 and S phase, was not uniform. A suspension of single nuclei was also more difficult to obtain. The nuclei tended to clump as they passed through the cell sorter, decreasing the purity of the sort due to minor disruptions of the flow. This will lead to increased contamination and therefore distorted ratios. A comparison between a lymphoblastoid and a fibroblastoid (adherent cell) flow sorter profile can be seen in Figure 5.28.

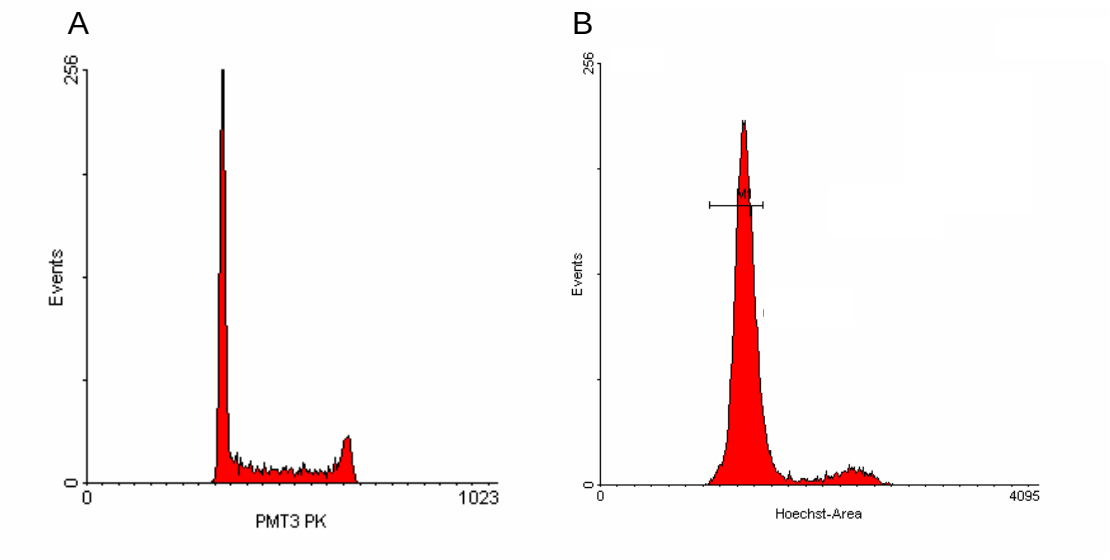


Figure 5.28: Comparison of the flow sort profiles of a lymphoblastoid cell line (A) and a fibroblastoid cell line (B).

Improving the discrimination between G1, S and G2 & M phase fractions obtained from other cell types will enable their replication timing to be assayed in the same way as achieved using the lymphoblastoid cell line. One way of achieving this would be to incorporate BrdU into nascent DNA. BrdU will therefore label cells in S phase only. Sorting nuclei using the combination of immunofluorescence against BrdU and propidium iodide staining for DNA content (Ormerod 2000) may be more effective at isolating nuclei in the S phase of the cell cycle for these cell types.

A further limitation of the method is the lack of heterochromatin and some other genome regions on the array. The clones on the array were selected from those used in the sequencing of the human genome (IHGSC 2001). Gaps in the draft sequence resulted in inevitable gaps in the replication timing profile of the chromosomes. Repeat regions of the genome are difficult to sequence leading to an under-representation of heterochromatic and centromeric DNA on the array. These sequences are late replicating (Gilbert 2002) and their lack of representation on the array will have an affect on the array normalisation. The array was normalised by the flow sort profile, which includes the representation of heterochromatin. An artefact of the normalisation process will therefore slightly bias measurements towards earlier replication.

The representation of repetitive heterochromatin sequences on the array in the appropriate amount would allow accurate normalisation of the array; however it would still be impossible to determine the exact replication timing of individual regions of highly repetitive heterochromatin using the current method. The effect of cross hybridisation of duplicated regions is detailed in 7.4.1. Repetitive heterochromatin represented on the array would cross hybridise with other repetitive sequences. The ratio reported by all repetitive regions would be an average replication timing of sequences that cross hybridise. The same limitations apply to regions of the genome with segmental duplications. Cross hybridisation results in the average replication timing of regions with similarity being reported. However it is unclear what effect the length and degree of homology will have on the replication timing ratio. As detailed in section 7.4.1, one way of circumnavigating this problem would be to use an array consisting of only unique sequence.

A related limitation of this method is the assay of regions where DNA replication is asynchronous, such as imprinted loci (Kawame, Gartler et al. 1995; Simon, Tenzen et al. 1999), the X chromosome in females (Avner and Heard 2001), immunoglobulin rearrangements and olfactory receptor genes (Goren and Cedar 2003). These are detailed in section 1.4.4. The flow sorting of the complete S phase ensures that both the early and late replicating alleles are hybridised to the same array. The ratio reported will be an average of the early and late replicating alleles and not the true replication time of either allele. To study the replication time of an imprinted region, cell lines with a uniparental disomy could possibly be used. S phase fractionation experiments such as those described in section 5.6 could also be used. However how regions involving immunoglobulin rearrangements and olfactory receptors would be assayed is unclear.

5.7.5: Verification of replication timing method

Replication timing has been previously assessed for a whole chromosome arm (Watanabe, Fujiyama et al. 2002), chromosome 11q, using flow sorting and real time PCR. This data was compared to the 1Mb resolution replication timing map of chromosome 11 obtained in section 5.5.1. The two replication timing profiles are very similar, with a correlation co-efficient of 0.69. This is despite the evaluation of

two different cell types and the total independence of the studies and methods used. This not only provides corroboration of our replication timing assay but also demonstrates the general similarity in the temporal programme of replication timing in these two different cell types.

Further verification was performed by selecting an early, two mid and a late replicating clone from the chromosome 22 replication timing profile described in section 5.3.1. and confirming the ratio of S:G1 by real time PCR. The two methods cannot be directly compared as the real time PCR can only assay a 150bp section of the genome. This is not comparable to the minimum 40Kb region sampled by the genomic arrays. A region of over 40Kb is likely to contain more than one replicon, whereas a region of just 150bp will not.

The standard deviation of the quadruplicates suggests the real time PCR is highly reproducible. The correlation between the replication time reported by arrays and that reported by real time PCR is strong ($r = 0.85$).

These two comparisons verify that genomic arrays can be used for the evaluation of replication timing. Unlike PCR based methods, the use of genomic arrays to assess replication ensures large regions of genome can be assessed at one time.

5.7.6: Assessment of replication timing using flow sorted S phase fractions.

Replication timing was assessed by detecting the increase in copy number within flow sorted fractions of S phase.

The majority of chromosome 22 loci were found to replicate in the first two fractions of S phase. Conversely most chromosome X loci replicated in the latter two fractions of S phase. This confirms what was found when the whole genome was analysed using complete S phase DNA hybridised against G1, with chromosome 22 replicating early and chromosome X replicating late.

In most cases the early replicating DNA increases copy number in the early S phase fractions. However, as shown in Figure 5.20 this did not reach the 2:1 ratio that would

signify that all replication was complete. 2:1 ratios were reached in the analysis of the G2/M fraction. This is not the case with late replicating fractions, which reached the 2:1 ratio in a linear fashion. It was observed that the late replicating clones exhibit a ratio of below 1:1 in the S1 and S2 fractions.

The scaling factors applied in Table 2.7 were calculated from the cell cycle profile. This means that the scaling factors applied are those that should be used for the whole genome. Application of an alternative normalisation factor, specific to chromosome 22 will increase the copy number change reported by the early replicating clones in the later S phase fractions to a ratio closer to 2:1. Contamination of the sorted S fractions by G1, G2 and other S DNA will also affect the ratios reported.

Loci reported by the single S:G1 phase hybridisation as early replicating display a copy number increase in the S1 and S2 fractions. This validates the single S:G1 hybridisation described in Chapter 5 for the assessment of replication timing.

Despite these limitations it is clear that genomic microarrays can be used to assess the replication timing of the majority of the genome. Verification by comparison with regions assayed by an alternative method shows that genomic arrays are highly effective at deducing replication timing over large regions of the genome.

5.7.7: Assessment of Replication Timing using High Resolution Arrays.

The construction of a high resolution array using 500bp PCR products is described in section 4.6. The data reported in section 3.4 from the test hybridisations displayed considerably more variation than that those reported for test hybridisations on other arrays. As described in section 4.6, the ratio reported by the PCR products from chromosome 22 reported more noise. It was therefore possible this noise in the ratios could be corrected. As a self:self hybridisation had been carried out, it was possible to divide the raw replication timing ratios by the self:self ratio reported for each sequence. The difference the inclusion of this analysis step makes is illustrated in Figure 5.29.

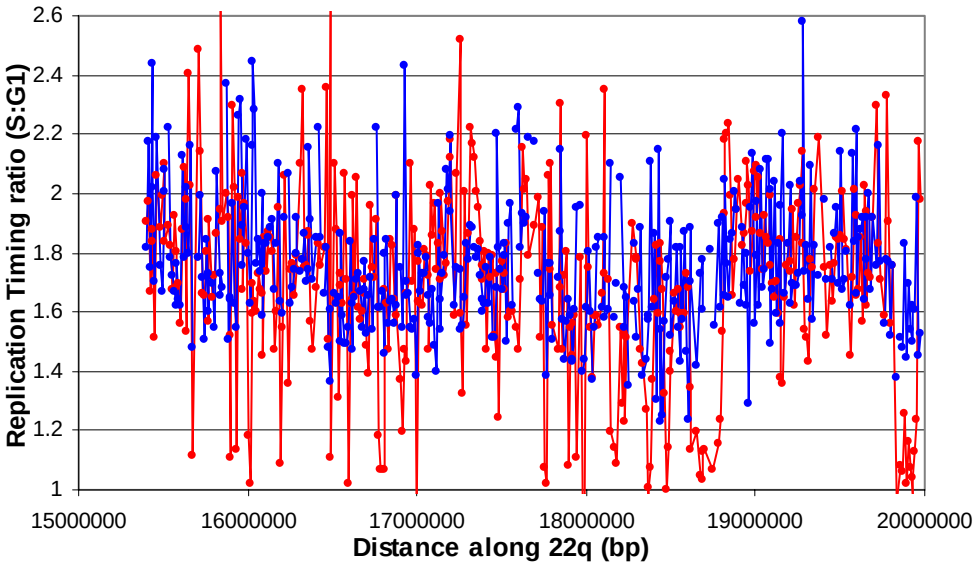


Figure 5.29: Correction against ratios reported for a self:self hybridisation on the high resolution array. Red: ratios reported prior to correction. Blue: ratios reported after correction as described.

Immediately it can be seen that the application of the additional correction factor tightens the ratios reported by the array from a spread of 0.84-2.75 and a standard deviation of 0.32, to a spread of 1.23-2.60 and a standard deviation of 0.22. However many points are still above the theoretical maximum of 2.0, reflecting inconsistencies with this method. Additional optimisation of the array, possibly by application of more accurate normalisation factors may increase the accuracy of the replication timing assay.

The analysis described in section 5.4.1 on the 10Kb resolution array shows that, in general the replication timing reported reflects that described by this region on the chromosome 22 tile path array. However this high resolution array does allow more detailed analysis. For example, a region 17.8-17.9Mb along chromosome 22 was reported as being early replicating by the 10Kb resolution array although the surrounding region was late replicating on the 22 tile path array. This 100Kb region is consistent with the size of one replicon, and so this result may identify a single early replicating replicon within a band of late replication. This indicates that although these arrays are currently not as accurate at reporting replication timing as the 22 tile

path array, they show the potential for detecting replication timing at a higher resolution.

The correlation between replication timing and GC content was also considered at high resolution. The GC content of each 500bp tile on the array was calculated and this was correlated with the replication time reported. The correlation at this level was weak with a correlation co-efficient of just 0.18, compared to a correlation coefficient of 0.27 when the same region is assayed using data from the tile path array. This shows a poorer correlation between replication timing and GC content when genomic DNA is assayed at this resolution.

The 500bp tile path resolution array showed increased variation compared to the tiling path arrays, although no ratios reported are above 2.2 or below 1. Additional peaks and troughs in the replication timing ratios were observed that were not apparent at the 10Kb resolution.

However, it can be argued that such marked peaks and troughs should not be seen at this high resolution. The array is made with 500bp overlapping PCR products. There are 275 PCR products on the array, covering 200Kb of sequence. As replicons are 40-100Kb in length the high resolution array is therefore thought to represent 2-5 replicons. The replication timing profile across the region does not reflect this. No groups of loci belonging to the same replicon, or replicon boundaries were identified. This could be due to the high coefficient of variation the consequence of which was that ratios within 0.4 of each other were not statistically different. This is very different to the 0.1 variation expected from the 22 tile path array, and would make it impossible to define replicons and replicon boundaries.

The high standard deviation of the method also makes it unfeasible to map replication origins using this method. DNA replicates at a rate of 50 base pairs per second. This means a replicon of 100Kb would take 33 minutes to replicate. A smaller 40Kb replicon would take 13 minutes to replicate. In the lymphoblastoid cell line assayed, S phase takes eight hours to complete. The tile path arrays have a standard error of 0.1 of S phase. This equates to 48 minutes. As this is greater than the time taken for a

replicon to replicate, variations within the replicon (such as the early replication of the origin) will not be identified.

Unfortunately time limitations did not allow full optimisation of the PCR product array. Despite these current limitations in assessing replication timing on the PCR product array it is clear from the work described in this Chapter that genomic clone arrays are highly effective at assessing replication timing and assaying large regions of the genome with a high accuracy.

5.7.8: Summary:

This Chapter has reported how microarrays have been used to assess human replication timing for the first time. Replication timing has been assayed on the whole genome at a 1Mb resolution, Chromosomes 1, 6, and 22 at a tile path resolution and a small region of chromosome 22 using 500bp PCR products.

Replication timing was then correlated with sequence features of the genome. Correlations with GC content, gene density and common repeat elements were observed.

The method was verified by comparison of the replication timing data produced for 11q with previously published data (Watanabe, Fujiyama et al. 2002). The replication timing of a selection of clones from chromosome 22 was assessed by real-time PCR. The replication timing of these clones was also found to correlate with the array method.