Human Molecular Genetics, 2004, Vol. 13, No. 2 ***191–202***
DOI: 10.1093/hmg/ddh016
Advance Access published on November 25, 2003

Replication timing of the human genome

Kathryn Woodfine[1], Heike Fiegler[1], David M. Beare[1], John E. Collins[1], Owen T. McCann[1], Bryan D. Young[2], Silvana Debernardi[2], Richard Mott[3], Ian Dunham[1] and Nigel P. Carter[1,*]

[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK, [2]Cancer Research UK, Molecular Oncology Group, Medical Oncology Unit, St Bartholomew's Hospital, London, UK and [3]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, UK

Received September 5, 2003; Revised October 30, 2003; Accepted November 9, 2003

We have developed a directly quantitative method utilizing genomic clone DNA microarrays to assess the replication timing of sequences during the S phase of the cell cycle. The genomic resolution of the replication timing measurements is limited only by the genomic clone size and density. We demonstrate the power of this approach by constructing a genome-wide map of replication timing in human lymphoblastoid cells using an array with clones spaced at 1 Mb intervals and a high-resolution replication timing map of 22q with an array utilizing overlapping sequencing tile path clones. We show a positive correlation, both genome-wide and at a high resolution, between replication timing and a range of genome parameters including GC content, gene density and transcriptional activity.

INTRODUCTION

It is widely held that replication of the human genome proceeds through a temporally ordered process which correlates with parameters related to gene activity, chromatin structure and nuclear position (1–3). However, the evidence supporting this view is restricted either in scale or scope. For example, replication banding of chromosomes has shown a correlation between R bands and gene density (1–3), but while this approach allows a genome-wide analysis, the fine detail of replication timing is lost. More detailed studies, utilizing fractionation of S phase followed by semi-quantitative PCR (4,5) or by the counting of FISH signals in S phase nuclei (6), show local variation in replication timing but due to the laborious nature of the methodology have been restricted to small regions at high resolution (5) or to single chromosome arms at a lower sampling resolution (4).

Genomic microarrays are commonly used to assess genome copy number differences in tumor DNA by comparative genomic hybridization [matrix-CGH (7), array-CGH (8)] with reference to normal diploid DNA. Our replication timing assay uses a similar approach to quantify the change in genomic copy number that occurs during S phase at each locus on the array. S phase DNA labeled in one color is hybridized together with unreplicated G1 phase DNA labeled in a second color onto genomic arrays. The fluorescence ratio for each sequence on the array is calculated as a measure of the time through S phase at which replication takes place (Fig. 1). This assay has many parallels with the assay developed by Selig *et al.* (6) using fluorescence *in situ* hybridization of clones to unsynchronized interphase nuclei.

RESULTS

The replication timing assay

The key elements of our approach require the separation of DNA from cells in S phase from the DNA of cells in G1 phase and the precision of genomic array hybridization to quantify relative DNA copy number changes to a high accuracy. Using an unsynchronized male lymphoblastoid cell line of normal karyotype, we separated S phase DNA from G1 phase DNA by flow sorting of cell nuclei stained with the DNA binding dye Hoechst 33258. After extraction of the DNA from sorted nuclei and differential labelling, the two fractions were hybridized simultaneously to the genomic clone array and the relative fluorescence of each measured. The ratio of S : G1 phase DNA (the replication timing ratio) reported by each sequence on the array after hybridization thus directly represented the average sequence copy number in the unsynchronized S phase fraction. Furthermore, because the proportion of nuclei in the unsynchronized S phase fraction in which any one particular sequence has replicated is proportional to the time at which this sequence replicates, the ratio also gave a measure of its replication time. Thus sequences with ratios close to 2 : 1 represent loci which replicate early in S phase as most of the

*To whom correspondence should be addressed. Email: npc@sanger.ac.uk

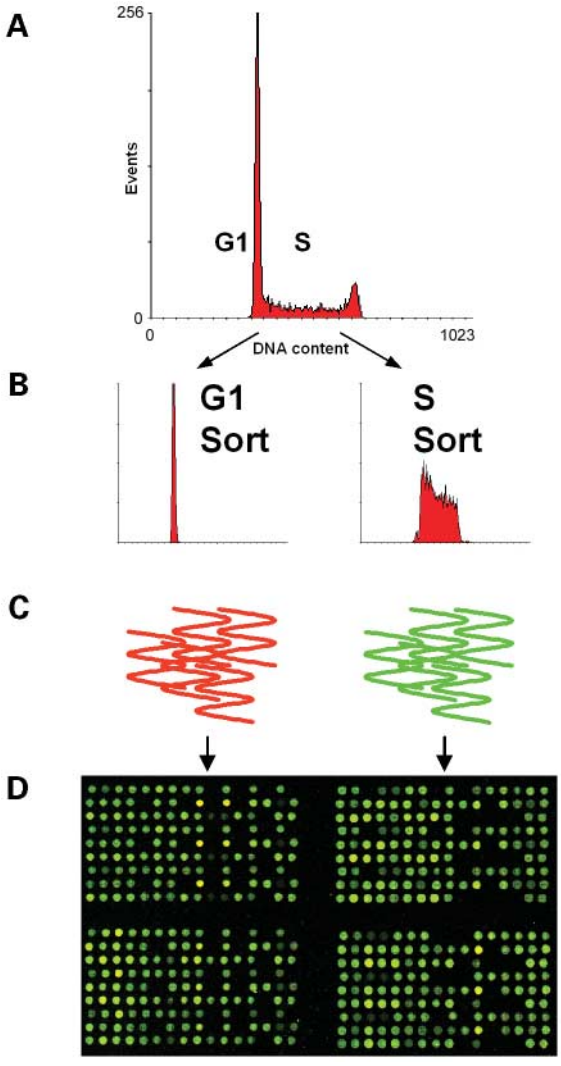

Figure 1. Experimental strategy for assessing genome wide replication timing using genomic microarrays. (**A**) Cell cycle profile of cycling human lymphoblastoid cell line after staining with Hoechst 33259. Cells in G1 and S phase of the cell cycle are sorted. (**B**) The sorted S and G1 phase fractions are checked for purity by re-analysis on the flow sorter and DNA extracted from the fractions. (**C**) The extracted DNA is differentially labeled with dCTP-Cy3 or dCTP-Cy5 using random primed labeling. (**D**) The labeled DNA is cohybridized to the array after preannealing with Cot1 DNA to suppress repeats. Early replicating sequences correspond to the spots with an increased S : G1 ratio.

nuclei will contain this replicated sequence. Conversely, loci with ratios close to 1 : 1 represent late replicating sequences.

This assay was used to quantify replication timing using two human genomic microarrays, the first covering the entire genome with clones (mean insert size 150 kb) spaced at ~1 Mb intervals (9) and a second higher resolution array covering the q arm of chromosome 22 with overlapping clones. Since we expect that ratios should only vary between 1 : 1 and 2 : 1, it was important to establish the accuracy and reproducibility of the assay. Therefore, we differentially labelled the same genomic DNA sample and hybridized it to the 22q array in a series of replicate experiments ($n = 5$). The coefficient of variation in the expected ratio of 1 : 1 for all clones was less than 3.5%. In addition, G1 fraction DNA versus itself and G2 DNA versus G1 DNA hybridizations were also performed on the 22q array. The average G1 : G1 ratio reported was 1.00 with a coefficient of variation of 5.35% and the average G2 : G1 ratio reported was 1.97 with a coefficient of variation of 5.25%. (Fig. S1, Supplementary Material). Finally the S : G1 phase DNA hybridizations used for evaluating replication timing (see below) were also performed in replicate on both the 1 Mb genome-wide array ($n = 4$) and the 22q array ($n = 4$). The coefficient of variation on both arrays was 5.5%. These results demonstrate the high level of accuracy and reproducibility afforded by genomic hybridization to DNA arrays.

Table 1. The mean replication timing ratio of each of the 24 human chromosomes. Chromosomes that replicate early have a higher ratio, close to 2. Late replicating chromosomes have a lower ratio, close to 1

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

Replication timing of the genome at 1 Mb resolution

The replication timing profiles for each chromosome obtained with the 1 Mb genome wide array are available for download at www.sanger.ac.uk/replication-timing.

Summary statistics are given in Table 1 and the replication timing profiles of two example chromosomes are shown in Figure 2.

In order to further validate our approach we were able to compare our results with an independent analysis of chromosome arm 11q previously published by Watanabe *et al.* (4). This group separated nuclei from a monocytic leukaemia cell line 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. In this way they could assign replication timing into categories approximating one-eighth of the time of the S phase. We were able to directly compare the replication timing profile of 11q

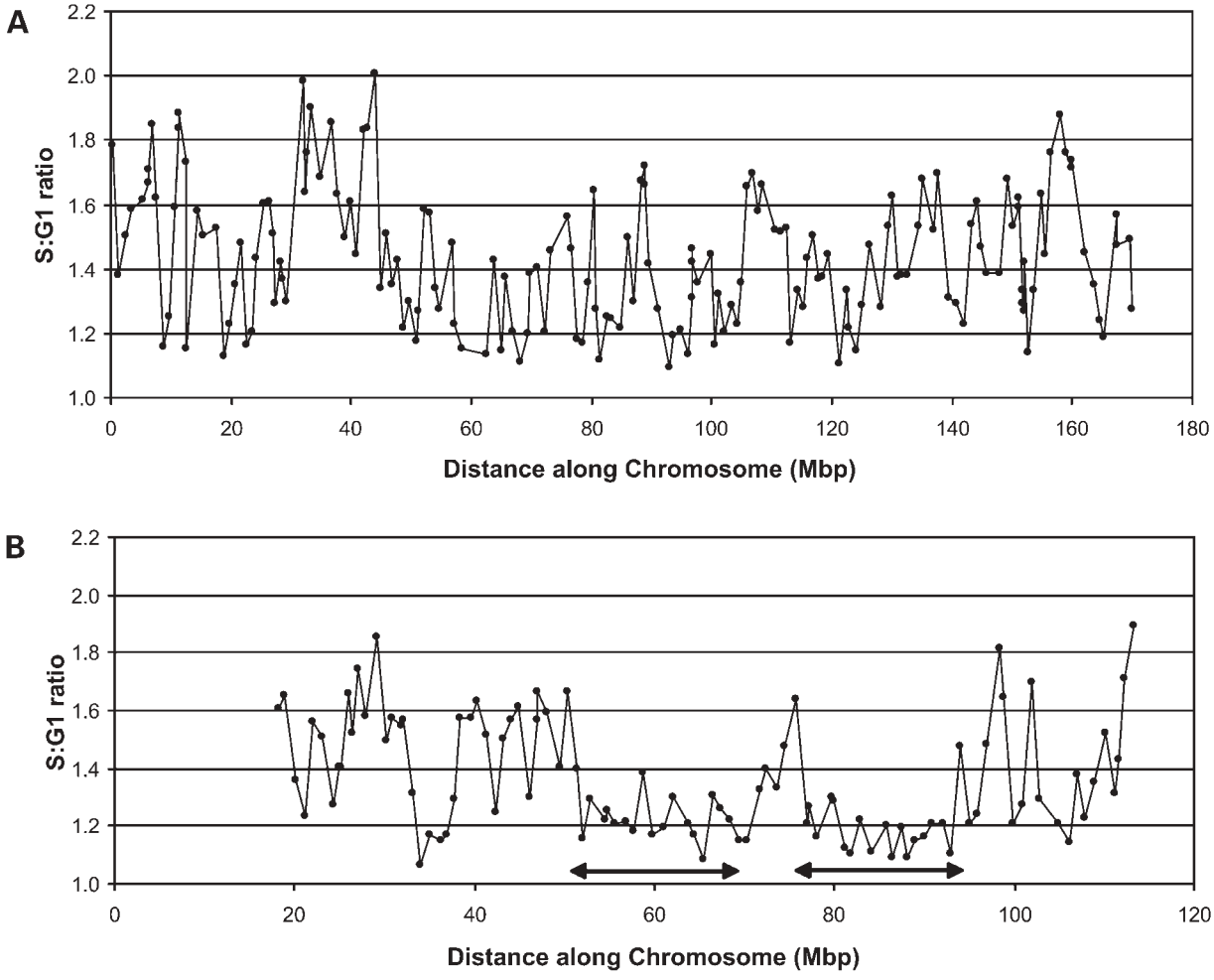

Figure 2. Replication timing profiles of two example chromosomes at a 1 Mb resolution. (**A**) The replication timing of chromosome 6 at a 1 Mb resolution. (**B**) The replication timing of chromosome 13 at a 1 Mb resolution. The black arrows indicate the gene deserts on chromosome 13. Data and graphs for all 24 chromosomes are available online for downloading at www.sanger.ac.uk/replication-timing.

obtained by Watanabe *et al.* (4) with our own data by remapping the positions of the STSs used onto NCBI Build 31 of the human genome (Fig. 3). We found a striking correlation ($r = 0.71$) between estimates of replication timing for regions of 11q represented on both the 1 Mb array and the PCR study (Fig. S2, Supplementary Material). It should be noted that this correlation was found despite the use of two different cell types, albeit both lymphoid in origin, and the total independence of the studies and the methods used. This not only provided validation for our replication timing assay but also demonstrates the general similarity in the temporal program of replication timing in these two different cell types.

The overall replication timing of entire chromosomes and chromosome arms has been related to the organization of chromosomes in the interphase nuclei, in particular three-dimensional chromosome position (10,11). Chromosomal condensation has also been linked to replication timing and the correct replication time may be a prerequisite for chromatin condensation patterns in the mitotic chromosome (12). Therefore correlation of replication time with other genome features at the whole chromosome level may help us understand the environment within chromosome territories. In the light of these relationships we examined the replication timing results across the whole genome in detail. Taken as a whole, chromosomes have mean replication times which are consistent with earlier analysis. For instance we were able to confirm that gene-rich chromosomes such as chromosomes 19 and 22 on average replicate early whilst gene-poor chromosomes such as 18 and 21 replicate late. Similarly chromosomes X and Y are generally late replicating. However the pattern of replication along the chromosome is not uniform, but is a mosaic of regions of early replicating DNA interspersed between late replicating regions.

Correlation with genomic features

To explain the patterns of replication timing observed across the genome we correlated our data with a range of genomic features (GC content of the sequence, gene density and *Alu* and LINE repeat sequence density) at both a whole chromosome level and clone by clone. First, looking at the whole chromosome level and performing regression analysis, we

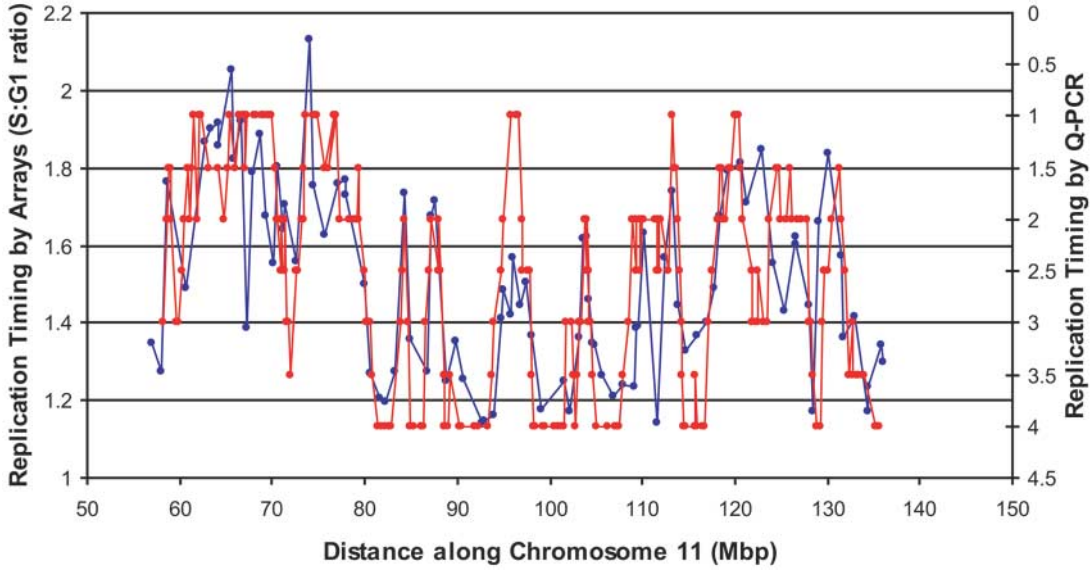

Figure 3. Comparison of replication timing data on 11q. Replication timing using genomic arrays (blue line, this study) and replication timing using semi-quantitative PCR [red line; Watanabe *et al.* (4)] plotted against position along the chromosome (NCBI 31). Replication timing data from Watanabe et al, 2002 was obtained from http://spinner.lab.nig.ac.jp/~tikemura/HumChr11ForHMG.html and was remapped onto the NCBI 31 genome build coordinates by electronic PCR (29).

found strong correlations between GC content of the chromosome and the mean replication timing for all chromosomes (Fig. 4). Significant positive correlations were also found between the mean replication timing and mean gene density and Alu repeat density and a negative correlation with LINE repeat sequence density, as shown in Table 2. This whole chromosome view approximates to earlier studies of R bands and confirms earlier correlations (13), although with more comprehensive knowledge of sequence content.

Next, to provide more detailed information on the local factors influencing replication timing, regression analysis was performed using the individual data points provided by each clone on the genome-wide array (Table 2). As well as the features studied on a chromosome level, the exon content of each clone was also calculated. The most significant correlation is still seen with GC content. GC-rich DNA is thus the best predictor of early replication. Noteworthy correlations are still seen with Alu repeats and measures of gene density, which are known to co-correlate with GC content. A negative correlation is seen between early replication and LINE-rich DNA. We also investigated how well a multiple regression that incorporated all the local factors performed at modeling replication timing. There was a 75% correlation between the 1 Mb array data and the fitted data, a small but highly statistically significant ($P < 10^{-16}$) improvement over the 70% correlation with GC content data alone (Table 2).

High resolution replication timing of chromosome 22q

The 1 Mb array gives sampling over the whole genome but in order to quantify patterns of replication timing at very high resolution a chromosome 22q sequence tile-path array was used. This array was constructed using a mixture of overlapping BAC, PAC, cosmid and fosmid clones which generated a replication timing data point on average every 78 kb across the sequenced portions of 22q. The replication timing profile of chromosome 22 is shown in Figure 5A. At this resolution we observed clear regions of similar replication timing covering several megabases of DNA, separated by adjacent regions of earlier or later replication timing with relatively sharp transition regions. Owing to the higher resolution of the 22 tile path array, many transitions in replication timing can be viewed that are not seen on the genomic array.

Chromosome 22 is very early replicating with most of the q arm replicating in the first half of the S phase. The most centromeric 9 Mb of the chromosome 22 sequence contains clones that include low copy repeats (LCRs) and areas of sequence that map to more than one region of the genome (14–16). Clones containing these sequences will report a ratio that is the average replication time of more than one region of the genome and therefore data from these clones were excluded from further analysis. The general correlations between replication timing and GC content, gene density and repeat sequence density observed across the whole genome and across whole chromosomes on the 1 Mb resolution array also exist in the detailed analysis along the length of 22q, although they are less strong with the exception of gene and exon density (Table 2, Fig. 5A–D, Fig. S2, Supplementary Material). Significant positive correlations between replication timing and intragenic DNA, exon density, Alu density and GC content were observed, although GC content and replication timing become less correlated at the distal end of 22q. A negative correlation between LINE density and replication timing was again seen. Multiple regression on the chromosome 22 replication timing data produced a correlation of 0.57 ($P < 10^{-16}$), a considerable improvement over the best single factor correlation (0.45 with Alu repeat density, Table 2).

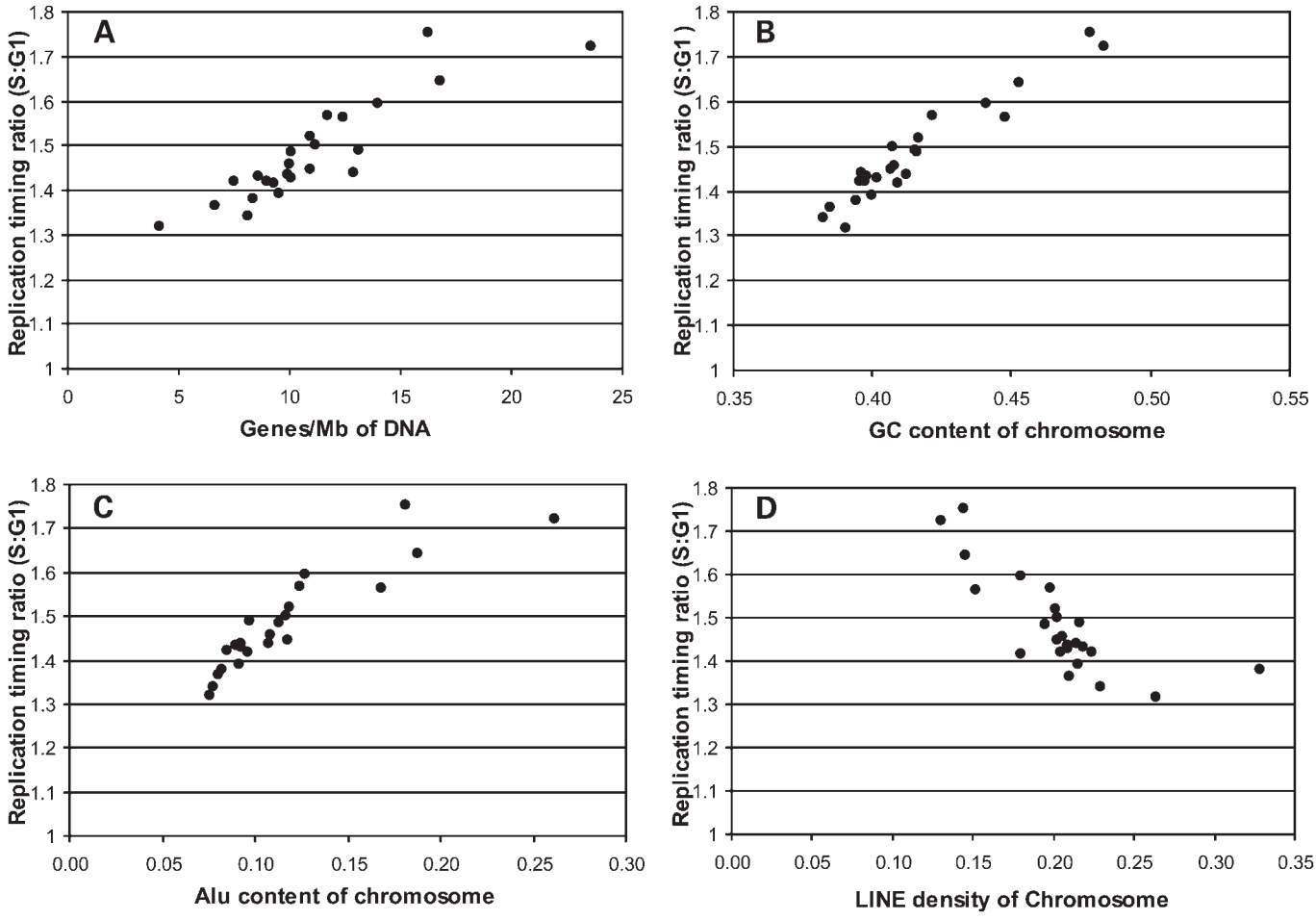

Figure 4. Correlation between replication timing and other features of the genome at a whole chromosome level. Each data point represents a chromosome. Statistically significant positive correlations were found between average replication timing of chromosomes and gene density ($y = 0.02x + 1.2$; **A**), GC content ($y = 0.07x - 0.17$; **B**), and Alu density ($y = 0.02x + 1.21$; **C**). Conversely a statistically significant negative correlation was found between average replication timing and LINE content of a chromosome ($y = -0.02x + 1.91$; **D**).

Table 2. Linear regression statistics correlating replication timing (RT) and other genome features on a chromosome-wide, 1 Mb and tile path resolution (for definitions, see Materials and Methods)

Genome feature	Correlation with RT on a chromosome-wide level	Correlation with RT at a 1 Mb resolution	Correlation with RT at a tile path resolution on chromosome 22
GC content	0.96	0.70	0.22
Alu repeat density	0.9	0.56	0.45
LINE repeat density	0.72	0.4	0.34
Gene density	0.89	0.35	0.41
Exon density	Not done	0.42	0.39
Combined analysis	Not done	0.76	0.57

Regions of coordinated replication

In order to assess the patterns of replication timing observed in the plot of the chromosome 22 data, we attempted to identify chromosome 22 regions of similar replication timing and regions which differed significantly in replication timing from adjacent stretches. A purpose-written perl program was used to find the optimal segmentation of the chromosome 22 data (see Materials and Methods). 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. Moreover, a permutation analysis indicates that the patterns of segmentation in the real data are highly non-random, with $P < 0.001$ for random rearrangement of the observed replication timing values along the chromosome producing as high a segmentation score. Results of this analysis

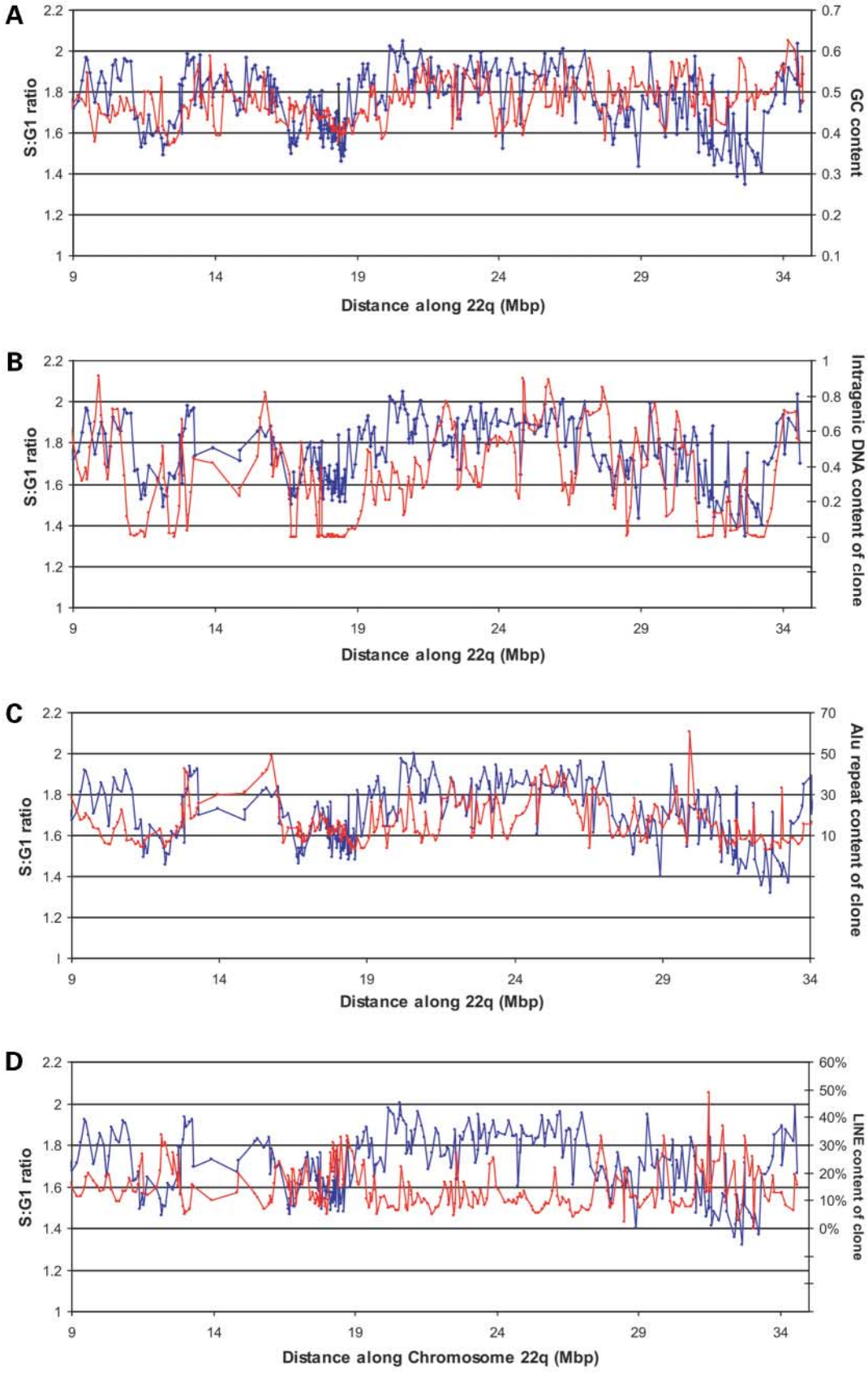

Figure 5. Correlation between replication timing and other features of the genome for chromosome 22q. (**A**) GC content. (**B**) Intragenic DNA. (**C**) Alu density. (**D**) LINE content.

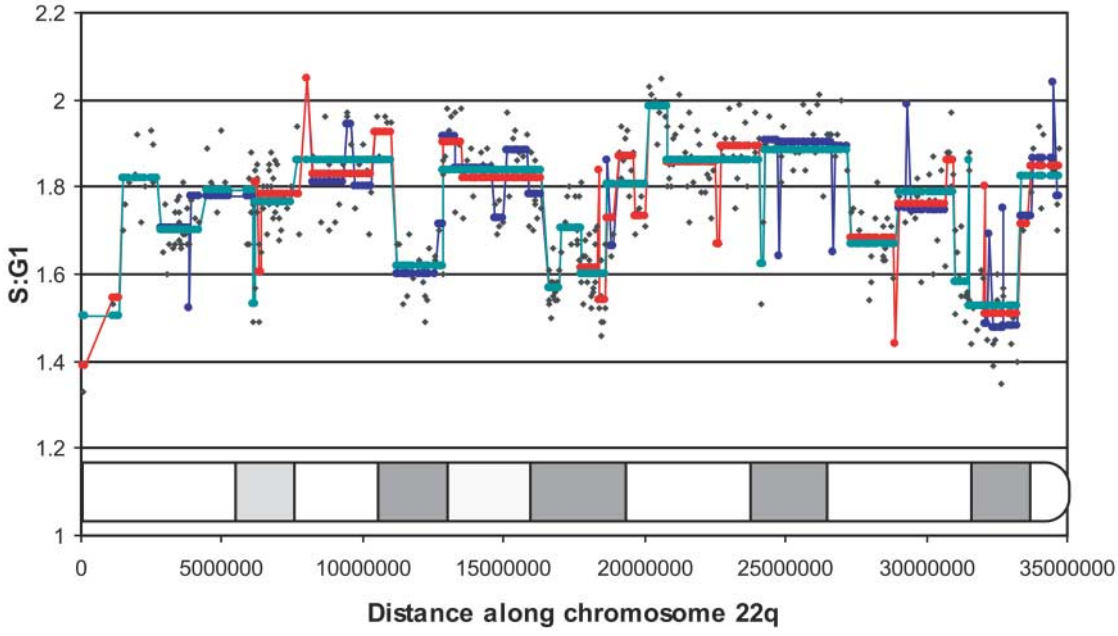

Figure 6. Analysis showing optimal segmentation of replication timing data across 22q. The graph shows the results of three runs of segmentation on the chromosome 22q data using representative segmentation penalty score (*B*) of 0.02 (blue), 0.04 (red) and 0.06 (green). Segmentation runs are plotted on top of the raw replication timing data (black circles). Regions of similar replication timing can then be compared to the banding pattern of chromosome 22 (redrawn from 24). See Materials and Methods for details.

for a series of representative values of *B* are shown in Figure 6. 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 where replication time is later than the genome average. The late-replicating regions correlate with particularly gene-poor regions of the chromosome. 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 160 kb apart (e.g. at ~11 100 000 bp and ~12 700 000 bp), suggesting disparate replication timing of adjacent replicons. If replication timing data can be acquired at higher density, this kind of approach could be used to define the borders of replicons.

Replication timing correlates with the probability of gene expression

Features such as GC content, Alu repeat density and gene density of a sequence all inter-correlate. Together or alone they may influence the replication timing of a region directly or they might act as markers of gene activity which also co-correlates, and which could be determining replication timing. However there has been some controversy concerning whether replication timing is correlated with gene expression. While no correlation was found in yeast (17), recently Schubeler *et al.* (18) reported a relationship between replication timing and gene expression in *Drosophila.* In order to determine whether such a correlation exists in human cells, the expression status of the genes represented within clones on the 1 Mb array was assessed. The Affymetrix U133A gene expression system was used, with mRNA prepared from the same logarithmically growing lymphoblastoid cell line used for replication analysis. Of the ~13 000 genes represented on the U133A array, 2063 genes were also represented within genomic clones on our 1 Mb genomic array. This analysis identified that 1013 of these genes were expressed in lymphoblastoid cells while the remaining 1050 genes did not show significant levels of expression. We then calculated the percentage of genes expressed per group of 50 genes ranked by their replication timing on the 1 Mb chip [cf. Schubeler *et al.* (18) for *Drosophila*]. Using logistic regression we found a significant positive correlation between the probability of gene expression and replication timing ($r = 0.61$, Fig. 7A). A stronger correlation was found when the same analysis was performed on data from the chromosome 22 overlapping clone array ($r = 0.92$ for windows of 50 genes and $r = 0.88$ for windows of 25 genes, Fig. 7C). However, we were unable to find a significant correlation with absolute levels of expression. An example replication timing and expression level profile for chromosome 2 is shown Figure 7B, illustrating this lack of correlation. Although in general, regions of high-level expression replicate early and regions containing transitions between early and late replication timing correlate with regions with a transition in gene expression levels, many disparate points were observed. It appears that whether the gene is expressed at all rather than the extent of expression is the important correlate with replication timing.

The rate of replication during the S phase

Using the data generated from the 1 Mb genome-wide arrays it was also possible to estimate the rate of replication. For this analysis the S phase was divided into centiles based on S:G1

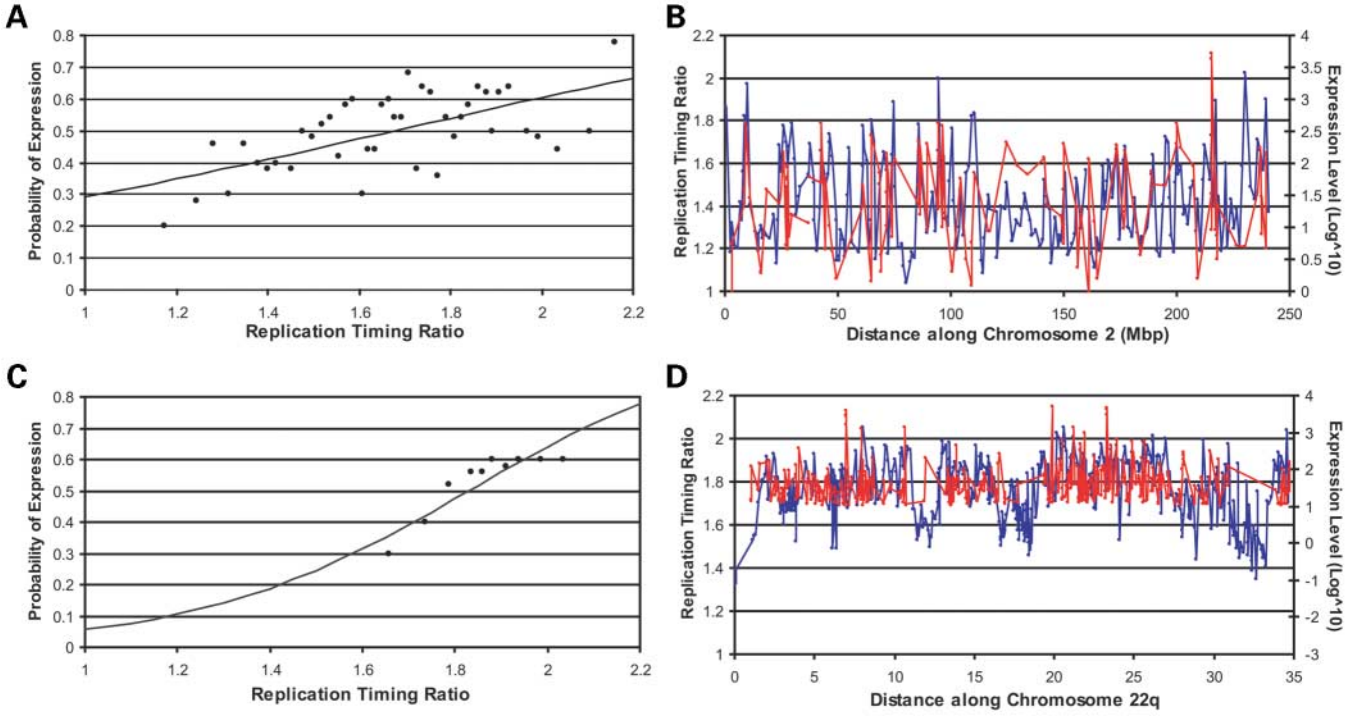

Figure 7. Relationship between transcription and replication timing. (**A**) Genes represented on the 1 Mb array were ordered and ranked into windows of 50, by replication timing and the probability of transcription was calculated. Logistic regression was then performed and the results plotted. (**B**) Expression level and replication timing plotted against chromosome position at a 1 Mb resolution on chromosome 2. (**C**) Genes represented on the chromosome 22 array were ordered and ranked into windows of 25, by replication timing and the probability of transcription was calculated. Logistic regression was performed and a strong positive correlation between replication timing and transcription observed. (**D**) Expression level and replication timing plotted against position on chromosome 22.

ratio. The number of loci replicating in each centile was calculated and the cumulative number of loci replicated was plotted against the proportion of the S phase completed (Fig. 8). This analysis suggests that replication commences slowly, but increases to a linear rate of replication at about a third of the way through the S phase. During this linear stage ~14% of the genome is replicated during each tenth of the S phase. The rate of replication then appears to slow as the end of the S phase is reached.

DISCUSSION

Using an approach based on genomic clone DNA microarrays we have been able to view the pattern of replication timing in human cells for the first time simultaneously across the whole genome at an ~1 Mb resolution. This approach allows us to build a map of the time through the S phase at which individual loci replicate at a whole genome level, providing a powerful tool for further studies of replication timing. Furthermore we have extended this approach to examine replication timing at ~70 kb resolution across a whole chromosome arm, giving an unprecedented view of the detailed patterns of replication timing. Microarrays have already been used to assess replication timing in yeast (17) and *Drosophila* (18). In *Drosophila*, replication timing was assessed using cDNA arrays. However our study is unique in using genomic clone microarrays for the assessment of replication timing in higher eukaryotes. By using genomic clones as the target, correlations can be drawn between replication timing and other features of the genome sequence not directly accessible with cDNA arrays.

Our replication timing assay is subject to some minor errors in the absolute measurement of the time at which replication takes place. The arrays were constructed from clones selected from the 'golden path' used in the sequencing of the human genome (19), so it is inevitable that gaps in the finished sequence lead to gaps in the replication timing profile of chromosomes assayed at a high resolution. In addition heterochromatin and centromeric DNA is under-represented on the genomic array. These sequences are known to be late replicating and, as they are missing from the arrays, an artefact of the normalization process will be a slight bias of all measurements towards later replication. There are also errors inherent in the separation of the S phase from the G1 phase of the cell cycle by flow cytometry. The sorting windows used to separate these two phases of the cell cycle are applied to the full cell cycle profile in which the G1 and S phases overlap slightly due to measurement variation. Applying curve fitting procedures to the cell cycle profile it is possible to extract best fit approximations to the distributions of the G1 and S phases (20,21). Using the cell cycle analysis program Cylchred (www.uwcm.ac.uk:10080/study/medicine/haematology/cytonetuk/documents/software.htm), we can estimate that within the S phase sorting window there are no contaminating G1 nuclei but within the G1 phase sorting window there are ~2% of nuclei in very early S phase. In addition, ~4% of the earliest S phase

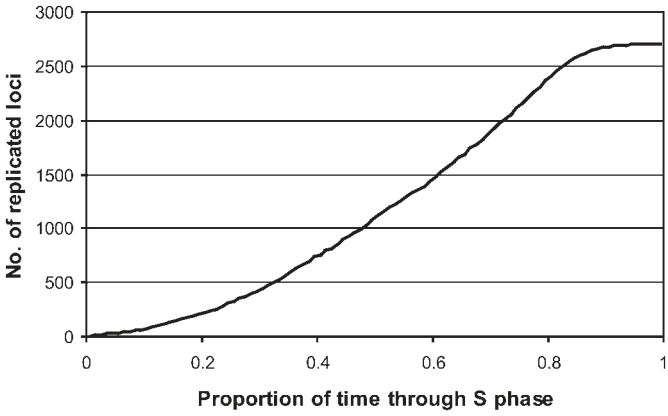

Figure 8. The rate of replication during the S phase of the cell cycle. S phase was divided into centiles based on the S : G1 ratio. The number of loci replicating in each centile was calculated and the cumulative number of loci replicated was plotted against the proportion of the S phase completed. Rate of replication is indicated by the slope of the curve plotted.

nuclei are not represented within the S phase fraction. The consequence of these enforced sorting inaccuracies is that, for the small number of the earliest replication loci, the theoretical replication timing ratio of 2.0 would be reduced to 1.93. The absolute value of the replication timing ratio is also highly dependent on the scaling factor applied to the normalized data which is estimated from the cell cycle profile as the median DNA content of the S phase fraction. This may well account for the small number of data points which exceed an absolute value of 2.0 in our data set. However, it should be noted that over 95% of all measurements are within a range representing a 2-fold increase in ratio and 98.6% of all measurements fall within a range representing a 2-fold increase plus and minus one standard deviation. Thus the main effect of an inaccuracy in determining the scaling factor is to offset the data points without affecting the relative distribution.

Our assay is also unable to correctly analyse regions of the genome that are imprinted. Imprinted loci are replicated asynchronously (22,23), but the ratio reported by our array will be a combination of the replication timing of the two alleles. Similarly data for recent segmental duplications where there is extensive sequence similarity will represent an average replication time for all the loci involved. In both cases information about the replication timing of individual loci will be lost.

It has long been acknowledged that mammalian R bands (GC-rich) replicate in the first half of S phase and G bands (GC-poor) replicate late (1–3). Our replication timing data demonstrates that these correlations persist at both a genome-wide level (on the 1 Mb array) and when a whole chromosome arm is studied at a high resolution. A comparison with 850 band resolution G banding as determined by average band length measurements (24) is shown in Figure S5 (Supplementary Material) for chromosomes 6 and 22. The replication timing profile at a 1 Mb level of chromosome 6 shows visual correspondence with G-banding in that G dark regions replicate late, such as those at 48–51 and 93–96 Mb along the chromosome, and G light bands replicate early, such as those 27–46 and 105–113 Mb along the chromosome. These comparisons are also maintained at tile-path resolution on chromosome 22. For example, the G dark bands between 16.6–18.8 and 112.2–12.5 Mb along the chromosome replicate late in comparison to surrounding regions while, conversely, G light bands located between 33.6–34.7 and 12.9–15.9 Mb along the chromosome replicate early. The correspondence between our replication timing measurements and G banding cannot be exact as the replication timing measurements are measured on a linear chromosome distance metric whereas the DNA content of G-light and G-dark bands varies due to the different levels of DNA condensation. In addition, the high resolution in replication timing that we can determine using our assay cannot be achieved even in the most highly banded chromosomes.

The direct correlation with GC level is the most statistically significant on a chromosome wide and genome-wide basis. A positive but less significant correlation was seen on the chromosome 22 tile path array. This difference persists if only clones with the same range of insert sizes from the chromosome 22 array as for the genome wide array are considered. One explanation for this difference may be that chromosome 22 is particularly GC-rich and the variation in GC content between the clones on the tile path array is not as great as that between the clones on the genome-wide array. Subtle changes in GC content at this resolution may not have a striking effect on replication timing. It is also particularly noticeable that GC content and replication timing become uncorrelated in 22q13 (30–34 Mb, Fig. 5A). This region is unusual for the human genome in that it is GC-rich but gene-poor. Replication timing is also late in this region suggesting that in this region it is gene content that is related to replication timing.

Although the correlations between GC content, gene density and replication timing are strong, there is also co-correlation with transcriptional activity. In *Drosophila* a correlation between replication timing and gene expression was observed (18) with genes replicating in the early S phase having a greater likelihood of expression than those replicating in the late S phase. Our analyses of human cells confirm that expressed genes tend to be replicated early in the S phase. Conversely, unexpressed genes usually replicate in the latter stages of S phase. Thus transcription and replication timing are closely linked. This is particularly clear at a higher resolution where, unlike gene density, the correlation with replication timing improves. It should be noted, however, that clones on the 1 Mb genome wide array which replicate late in the S phase are under-represented on the Affymetrix U133A array (Fig. 7D). This is because few genes are found in this region and so late replication is associated with regions of gene sparseness. It is also clear that the level of transcription is not important.

From our genome-wide replication timing data we were able to estimate and model the rate of replication throughout S phase. Replication appears to start slowly, but increases to a linear rate of replication at about a third of the way through the S phase, finally again appearing to slow as the end of the S phase. The slow initial rate of replication is supported by the shape of the distribution of the S phase as measured on the flow cytometer (see the S phase sorted fraction in Fig. 1) 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 the 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. However, the slow rate of replication at the end of the 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 the S phase onwards. As most heterochromatin will replicate during this late stage, and heterochromatic regions are not represented on the 1 Mb genome-wide arrays, the rate of replication for this final part of the S phase is likely to be underestimated.

Models based on open chromatin (25) suggest that transcriptional activity enables early replication by promoting access of the replication machinery to the DNA (26). Under these models, regions of the genome that are transcriptionally silent because of their lack of genes would be expected to replicate late in the S phase and we are able to confirm predictions for regions of the genome such as the gene deserts on chromosomes 13 and 14. In addition the gene-poor but GC-rich distal end of chromosome 22 replicates late, lending further support to the hypothesis that transcriptionally active chromatin and early replication are co-driven. Since GC content, gene density and transcriptional activity all co-correlate strongly with replication timing, it is not yet possible to unravel the causative relationship behind these intercorrelations. However high-resolution study of exceptional regions of the sort found in 22q13 could provide clues to unraveling the primary driver of replication time. Finally further studies on additional tissues will highlight how changes in expression or epigenetic modifications may affect replication timing. Understanding replication timing on a genome-wide basis will provide important insights into the regulation of DNA replication.

MATERIALS AND METHODS

Tissue culture

A human male lymphoblastoid cell line with a normal (46, XY) karyotype (CO202 ECCAC no. 94060845) was cultured in RPMI media supplemented with 16% FCS, 2 mM L-glutamine, 100 units/ml penicillin and 100 mg/ml streptomycin. Cells were harvested 26 h after subculture to maximize the percentage of nuclei in the S phase. The cells were centrifuged at 300*g* for 5 min, the pellet resuspended in 75 mM KCl and incubated at room temperature for 15 min. After further centrifugation the pellet was finally resuspended in polyamine buffer (80 mM KCl, 20 mM NaCl, 2 mM EDTA, 0.5 mM EGTA, 15 mM Tris, 3 mM dithiothreitol and 0.25% vol/vol Triton X-100, pH 7.2) at a concentration of $\sim 6 \times 10^6$/ml, stained with 2 μg/ml Hoechst 33258 and sorted immediately.

Flow sorting and DNA precipitation

Stained nuclei were separated using a Coulter Elite ESP flow sorter (Beckman-Coulter, Fullerton, CA, USA) into S and G1 phase fractions in sheath buffer comprising 0.1 M NaCl, 0.01 M Tris pH 7.4 and 0.001 M EDTA. EDTA (250 mM), sodium lauroyl sarcosine (1%) and proteinase K (200 μg/ml) were added, and the nuclei were incubated overnight at 42°C. After addition of PSMF to a final concentration of 4 mg/ml and incubation at room temperature for 40 min, DNA was precipitated by the addition of NaCl (final concentration of 640 mM) and 2 vols 100% ethanol and resuspended in 10 mM Tris pH 7.4, 0.1 mM EDTA buffer at a concentration of ~500 mg/ml.

DNA labeling

S and G1 phase DNA was differentially labeled as described (9). Briefly, 450 ng of the S phase DNA was labeled with dCTP-Cy3 and 450 ng of G1 phase DNA was labeled with dCTP-Cy5 using a Bioprime Labeling kit (Invitrogen, Carlsbad, CA, USA). After labeling, the DNA was purified using a G50 spin column (Amersham Pharmacia, Buckinghamshire, UK).

Microarray preparation

For the experiments requiring a 1 Mb resolution across all 24 chromosomes, we used the genomic array as described (9). For high-resolution studies on chromosome 22, a tile path array was constructed containing 447 clones, 444 of which represented golden path sequencing clones (16) while the other three clones required to complete the tiling path were identified by their BAC end sequence positions. Clones were tested to check that they were bacteriophage negative and verified by *Hind*III fingerprinting or STS (sequence tagged site) PCR. After all verification steps there were two gaps in the chromosome 22 sequence; between accession numbers AP000526–AC005529 (1165 kb) and AC005500–AP000555 (671 kb) The 22q genomic array was constructed as previously described (9) with clones spotted in triplicate. The final grid size was 6 cm^2.

Hybridization and array analysis

Hybridizations were carried out as described (9). The arrays were scanned using an Axon 4000B scanner (Axon Instruments, Burlingame, CA, USA) and images quantified using 'Spot' software (27). For the 1 Mb resolution array, raw fluorescence ratios were normalized by dividing each ratio by the mean ratio of all clones and scaled by a value representing the median DNA content of the S phase fraction calculated from the cell cycle histogram. For the 1 Mb array the median DNA content of the S phase fraction was 1.44. For the chromosome 22 tiling path array, normalization was carried out in a similar way except the scaling factor used was calculated from the mean replication timing ratio reported for chromosome 22 on the 1 Mb resolution array (1.75).

Expression analysis

Total RNA was extracted from lymphoblastoid cells using the Trizol (Gibco-BRL, Cheshire, UK) purification method. First- and second-strand cDNA synthesis was performed from 10 μg of total RNA, with the Superscript ds-cDNA Synthesis Kit (Gibco-BRL, Cheshire, UK), using 100 pmol of a HPLC purified T7-(T)24 primer. Amplified biotinylated complementary RNA was then produced with an *in vitro* transcription labeling reaction, performed according to the manufacturer's

recommendation (Enzo Diagnostics, Farmingdale, NY, USA). Samples with a yield greater than 40 µg of cRNA were subsequently hybridized to an Affymetrix U133 oligonucleotide arrays (Affymetrix Santa Clara, CA, USA). Hybridization was performed at 45°C for 16 h.

Arrays were washed and stained with streptavidin-phycoerythrin (SAPE, Molecular Probes, Leiden, The Netherlands). Signal amplification was performed using a biotinylated anti-streptavidin antibody (Vector Laboratories Burlingame, CA, USA) following the recommended Affymetrix protocol for high density chips. Scans were carried out on a GeneArray scanner (Agilent Technologies Palo Alto, CA, USA). The fluorescence intensities of scanned arrays were analysed with Affymetrix GeneChip software. The Affymetrix Microarray Suite 5.0 was used for the quantification of gene expression levels. Global scaling was applied to the data to adjust the average recorded intensity to a target intensity of 100. Quantification data was exported from Affymetrix Microarray Suite 5.0 into Excel for further analysis. Presence or absence of gene expression was determined by a 'present' call in any of the oligos representing a gene, as determined by Affymetrix Microarray Suite 5.0.

Genome parameters

For 1 Mb and chromosome 22 analysis genome parameters were calculated over individual clones. Where available end sequence information was used to calculate the 'clone window' defined by the true clone boundary co-ordinates and clone length, otherwise the accessioned sequence was used. Details of clones used on the 1 Mb chip is available in the Ensembl genome browser Cytoview pages www.ensembl.org/Homo_sapiens/cytoview and the average sequence length of the clones was 128 kb, parameters were extracted from Ensembl data. On chromosome 22 genome parameters were extracted from the current annotation (28). Information used to calculate the genome parameters (gene density, GC content, *Alu* density and LINE density) averaged over the whole chromosome were obtained from the University of California, Santa Cruz Website http://genome.ucsc.edu. All coordinates for genome-wide microarray analysis are based on NCBI Build 31 of the human genome sequence. For chromosome 22 tile-path array analyses coordinates are mapped to the 22q sequence build described in Collins *et al.* (28). To remap these coordinates to NCBI 31 coordinates add 13 Mb to adjust for the unsequenced part of 22p-cen.

GC content was defined as the fraction of G+C bases found in the clone sequence, or average G+C content of the chromosome. Gene density was defined as the fraction of intronic and exonic DNA found in the clone sequence, or the average number of genes per megabase over a chromosome (number of Genes/chromosome length). Exon density was defined as the fraction of exonic DNA found in the clone sequence. Alu density was the defined as fraction of clone or chromosome sequence involved in Alu repeats. LINE density was the fraction of clone or chromosome sequence involved in LINE repeats.

Data segmentation analysis

A purpose-written perl program, available from the authors, was used to find the optimal segmentation of the replication timing (RT) data. Suppose a chromosome contains n RT signals arranged in genome order. Within each segment, starting at coordinate i and ending at coordinate j, we define the score S_{ij} equal to the sum of squared deviations of the RT values from the mean RT signal μ_{ij} for the segment. The optimal segmentation pattern (i.e. the number of segments and coordinates of segment boundaries) is chosen which minimizes a function, W_n, based on the sum of segment scores plus a penalty score B for each segment transition. Let W_k be the score of the optimal segmentation for coordinates 1 through k. Then $W_0 = 0$ and $W_k = \min_{i<k} \{W_{i-1} + B + S_{ik}\}$ for all $k > 0$. The degree of segmentation is controlled by the value of B. The optimal segmentation is found by backtracking from the terminal value W_n. The statistical significance of W was determined by re-running the program on 1000 permuted data sets in which the order of observed RT signals was shuffled. The P-value for the test of the null hypothesis that the observed segmentation score could have arisen by chance is estimated as the proportion of times the permuted W score exceeded the observed score.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online and at www.sanger.ac.uk/replication-timing/.

ACKNOWLEDGEMENTS

We would like to thank Tony Cox of the Ensembl team and Jim Kent at UCSC for help in calculating the genome parameters for the 1 Mb set and the whole chromosome, Carol Scott for help in choosing the chromosome 22 clones and Lisa French, Paul Hunt, Carol Carder and Sean Humphray in the Sanger Institute Core Mapping Facility for their help constructing the clone sets and Cordelia Langford, the Sanger Institute Microarray Facility and Dave Vetrie for development of the surface chemistry and for array printing. This work was supported by the Wellcome Trust. K.W. is supported by the Medical Research Council.

REFERENCES

1. Dutrillaux, B., Couturier, J., Richer, C.L. and Viegas-Pequignot, E. (1976) Sequence of DNA replication in 277 R and Q-bands of human chromosomes using a BrdU treatment. *Chromosoma*, **58**, 51–61.
2. Ganner, E. and Evans, H.J. (1971) The relationship between patterns of DNA replication and of quinacrine fluorescence in the human chromosome complement. *Chromosoma*, **35**, 326–341.
3. Holmquist, G., Gray, M., Porter, T. and Jordan, J. (1982) Characterization of Giemsa dark- and light-band DNA. *Cell*, **31**, 121–129.
4. Watanabe, Y., Fujiyama, A., Ichiba, Y., Hattori, M., Yada, T., Sakaki, Y. and Ikemura, T. (2002) Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions. *Hum. Mol. Genet.*, **11**, 13–21.
5. Sinnett, D., Flint, A. and Lalande, M. (1993) Determination of DNA replication kinetics in synchronized human cells using a PCR-based assay. *Nucl. Acids Res.*, **21**, 3227–3232.

6. Selig, S., Okumura, K., Ward, D.C. and Cedar, H. (1992) Delineation of DNA replication time zones by fluorescence in situ hybridization. *EMBO J.*, **11**, 1217–1225.
7. Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Dohner, H., Cremer, T. and Lichter, P. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**, 399–407.
8. Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y. *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.
9. Fiegler, H., Carr, P., Douglas, E.J., Burford, D.C., Hunt, S., Smith, J., Vetrie, D., Gorman, P., Tomlinson, I.P. and Carter, N.P. (2003) DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosomes Cancer*, **36**, 361–374.
10. Cross, S.H., Clark, V.H., Simmen, M.W., Bickmore, W.A., Maroon, H., Langford, C.F., Carter, N.P. and Bird, A.P. (2000) CpG island libraries from human chromosomes 18 and 22: landmarks for novel genes. *Mamm. Genome*, **11**, 373–383.
11. Zink, D., Bornfleth, H., Visser, A., Cremer, C. and Cremer, T. (1999) Organization of early and late replicating DNA in human chromosome territories. *Exp. Cell Res.*, **247**, 176–188.
12. Gerbi, S.A. and Bielinsky, A.K. (2002) DNA replication and chromatin. *Curr. Opin. Genet. Dev.*, **12**, 243–248.
13. Cohen, S.M., Cobb, E.R., Cordeiro-Stone, M. and Kaufman, D.G. (1998) Identification of chromosomal bands replicating early in the S phase of normal human fibroblasts. *Exp. Cell Res.*, **245**, 321–329.
14. Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W. and Eichler, E.E. (2002) Recent segmental duplications in the human genome. *Science*, **297**, 1003–1007.
15. Bailey, J.A., Yavor, A.M., Viggiano, L., Misceo, D., Horvath, J.E., Archidiacono, N., Schwartz, S., Rocchi, M. and Eichler, E.E. (2002) Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am. J. Hum. Genet.*, **70**, 83–100.
16. Dunham, I., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., Ainscough, R., Almeida, J.P., Babbage, A. *et al.* (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489–495.
17. Raghuraman, M.K., Winzeler, E.A., Collingwood, D., Hunt, S., Wodicka, L., Conway, A., Lockhart, D.J., Davis, R.W., Brewer, B.J. and Fangman, W.L. (2001) Replication dynamics of the yeast genome. *Science*, **294**, 115–121.
18. Schubeler, D., Scalzo, D., Kooperberg, C., van Steensel, B., Delrow, J. and Groudine, M. (2002) Genome-wide DNA replication profile for *Drosophila melanogaster*: a link between transcription and replication timing. *Nat. Genet.*, **32**, 438–442.
19. International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
20. Watson, J.V., Chambers, S.H. and Smith, P.J. (1987) A pragmatic approach to the analysis of DNA histograms with a definable G1 peak. *Cytometry*, **8**, 1–8.
21. Ormerod, M.G., Payne, A.W. and Watson, J.V. (1987) Improved program for the analysis of DNA histograms. *Cytometry*, **8**, 637–641.
22. Simon, I., Tenzen, T., Reubinoff, B.E., Hillman, D., McCarrey, J.R. and Cedar, H. (1999) Asynchronous replication of imprinted genes is established in the gametes and maintained during development. *Nature*, **401**, 929–932.
23. Kawame, H., Gartler, S.M. and Hansen, R.S. (1995) Allele-specific replication timing in imprinted domains: absence of asynchrony at several loci. *Hum. Mol. Genet.*, **4**, 2287–2293.
24. Francke, U. (1994) Digitized and differentially shaded human chromosome ideograms for genomic applications. *Cytogenet. Cell. Genet.*, **65**, 206–218.
25. Gilbert, D.M. (2002) Replication timing and transcriptional control: beyond cause and effect. *Curr. Opin. Cell Biol.*, **14**, 377–383.
26. McCune, H.J. and Donaldson, A.D. (2003) DNA replication: telling time with microarrays. *Genome Biol.*, **4**, 204.
27. Jain, A.N., Tokuyasu, T.A., Snijders, A.M., Segraves, R., Albertson, D.G. and Pinkel, D. (2002) Fully automatic quantification of microarray image data. *Genome Res.*, **12**, 325–332.
28. Collins, J.E., Goward, M.E., Cole, C.G., Smink, L.J., Huckle, E.J., Knowles, S., Bye, J.M., Beare, D.M. and Dunham, I. (2003) Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res.*, **13**, 27–36.
29. Schuler, G.D. (1998) Electronic PCR: bridging the gap between genome mapping and genome sequencing. *Trends Biotechnol.*, **16**, 456–459.