Evidence that both G + C rich and G + C poor isochores are replicated early and late in the cell cycle

Article (Published Version)

Nucleic Acids Research, Vol. 20, No. 7 1497–1501

Evidence that both G+C rich and G+C poor isochores are replicated early and late in the cell cycle

Adam Eyre-Walker
Institute of Cell Animal and Population Biology, University of Edinburgh, EH9 1JT, UK

Received January 31, 1992, Revised and Accepted March 12, 1992

ABSTRACT

Since the G+C content of a gene is correlated to that of the isochore in which it resides, and early replicating isochores are thought to be relatively G+C rich, early replicating genes should also be rich in G+C. This hypothesis is tested on a sample of 44 mammalian genes for which replication time data and sequence information are available. Early replicating genes do not appear to be more G+C rich than late replicating genes, instead there is considerable variation in the G+C content of genes replicated during both halves of S phase. These results show that both G+C rich and poor fractions of the genome are replicated early and late in the cell cycle, and suggest that isochores are not maintained by the replication of DNA sequences in compositionally biased free nucleotide pools.

INTRODUCTION

A paradox seems to have gone unnoticed. It is believed that G+C rich isochores and housekeeping genes replicate early in the cell cycle, with G+C poor isochores and some tissue specific genes replicating late (1–3). Since the G+C content of a gene is correlated to the isochore in which it lies (4,5) housekeeping genes should be G+C rich compared to tissue specific sequences. However there appears to be no difference in the G+C contents of housekeeping and tissue specific genes (6).

The link in this paradox I wish to focus on is the early replication of G+C rich isochores, since the evidence for it can be interpreted in two ways. Evidence for the early replication of G+C rich DNA comes from the 3–5% difference in G+C content that has been measured between the early and late replicating fractions of the genome (7–11) and the coincidence of chromosome bands produced by G+C content sensitive methods, such as quinacrine staining, and replication time bands (1,12). The simplest and most popular interpretation of these observations is that most, if not all of the isochores replicated early in the cell cycle have a higher G+C content than those replicated late in S phase (3,12–14). However the observations are equally consistent with the replication of all fractions of the genome both early and late in the cell cycle, with the early replicating DNA only being on average slightly more G+C rich. The difference is very important since it has implications for our understanding of chromosome structure and evolution; in particular how isochores are maintained. It also seems inappropiate to talk about early replicating DNA being more G+C rich if in fact most of the variation in G+C content is within the early and late replicating fractions, not between them.

In order to distinguish between these alternatives a set of genes for which there are replication time data and sequence information was compiled. If we assume that gene and isochore G+C contents are correlated the G+C contents of the genes will give an insight into the range of isochore G+C contents replicated during the two halves of S phase. The data will also allow us to eliminate the possibilities that the paradox has arisen, (i) because many tissue specific genes replicate early in the cell cycle, and (ii) because the relationship between isochore and gene G+C contents is different for housekeeping and tissue specific sequences.

MATERIALS AND METHODS

Replication Time Data

Data on the replication time of specific genes was taken from Holmquist (3) with minor modifications (see below). His list is a compilation of data from references 15–20. These studies suggest that all genes expressed in a tissue are replicated early, but that unexpressed genes may replicate at any time during S phase. If a gene does replicate late in one cell type it does so in most other cell lines in which it is not expressed. Therefore genes which were found to replicate late in most cell types, in which they were not expressed, were classified as late replicating. All other genes were classed as early replicating. The classification was the same as that given by (3) except for albumin, which appears to have been misclassified, and complement C4 which was not included in the compilation. It is worth pointing out that the replication time of most genes was coincident over several cell types from several species. In particular there were no genes with different replication times between the two species groups (rodents and primates).

Sequence Information

Sequence information was taken from the Genbank (Release 68) and Embl (Release 27) databases using the GCG sequence analysis package (20). Accession numbers are available on request. Human and mouse sequences were extracted for all genes for which replication time data were available. If a mouse gene was not available the rat sequence was used instead. The G+C contents of mouse and rat genes are very similar (21,22) so mixing rat and mouse genes should not lead to substantial bias.

If human sequences were not available other primate sequences were used. The classification as to housekeeping or tissue specific expression was taken from Holmquist (3) who gave no details as to how the classification was performed.

Our primary interest in the G+C contents of early and late replicating genes is not the compositions of the sequences themselves, but what they tell us about the isochore in which they reside; i.e we are interested in whether G+C rich and G+C poor isochores replicate both early and late. It is therefore important to ensure that a particular isochore is only represented once in the data set, by including only one gene from a set of linked, or recently diverged, genes. Since isochores are thought to be at least 300kb in length (12) genes within this distance of one another were regarded as representing the same isochore. The average G+C content over a set of linked genes was not used because it is possible for a linkage group to traverse two or more isochores of different compositions. Instead the longest sequence was used. Small scale physical distance information (<300kb) was taken from references 17 and 19. Large scale linkage was also checked using HGM10.5 (23,24) and a mouse map (25). No genes were excluded on the basis of this information because of the scale and the lack of accuracy involved. However suffice it to say that only six genes were found to be 'linked' (within a centimorgan or in the same chromosome band). Beta-globin and c-Ha-ras map to the same human chromosome band, 11p15.5, but are some 18 centimorgans apart in mice; immunoglobulin kappa constant and variable map to the same chromosome band, 2p12, in humans and the same centimorgan in mice, 6.32; and arginine succinate synthetase and c-abl are the same distance along chromosome 2 in mice.

If several sequences from a dispersed multigene family were available with replication time information (e.g beta and gamma actin), only one sequence was used since any recently diverged members will tend to correlate with the G+C content of the 'parental' isochore, not that of their present location. Such sequences will therefore tend to contribute information about the same isochore.

One further problem with multigene families is identifying which member the replication time is actually known for, especially if some members of the family are quite disimilar to each other. For instance the rodent and primate placental lactogen genes have very different G+C contents which suggests that they are paralogous, and such paralogous genes could have different replication times. However this source of error is only relevent if the paralogous genes have different G+C contents, of which there is little evidence in the data set (table 1) and most of the sequences used are probably single copy genes. Therefore any errors should be small.

Testing The Data

Differences in the distribution of G+C contents, say of early and late replicating genes, were tested with a Mann-Whitney test. This tests whether two sets of data could have come from the same distribution, and fails if the medians are different, or if the medians are the same but the shapes of the distributions are asymmetrical and different.

Such tests ask whether two sets of data could have come from the same distribution, whereas we want to ultimately ask a slightly more subtle question: are the gene G+C contents we observe consistent with all the early replicating isochores being more G+C rich than the late replicating isochores? In order to do this we need to take into account the less than perfect correlation between gene and isochore G+C contents; i.e it is possible for all early replicating isochores to be more G+C rich than late replicating isochores and yet for there still to be some overlap in the G+C contents of early and late replicating genes. The approach taken was as follows: isochore G+C contents were randomly generated from gene G+C contents in a way consistent

Table 1. The expression status, replication time and G+C content of a set of primate and rodent genes.

Gene	Rep Time[a]	Exp[b]	Time known[c]	Primate 3[d]	Time known[c]	Rodent 3[d]
HPRT	E	H	√	39.6	√	41.5
APRT	E	H	√	81.6	√	74.3
CAD	E	H	√	71.5	√	NA
DHFR	E	H	√	42.5	√	47.8
Argininesuccinate synthetase	E	H	√	74.7	√	67.9
Glucose-6-phosphate dehydrogenase	E	H	√	85.1		62.5
β-tubulins	E	H	√	81.8	√	71.8
Phosphoglycerate kinase 1	E	H	√	55.8		54.3
Tyrosine aminotransferase	E	H	√	NA	√	62.7
β-actin	E	H	√	84.5		73.0
Metallothionein I	E	H	√	80.0	√	88.3
c-myc	E	H	√	76.7	√	75.8
c-Ha-ras	E	H	√	81.4	√	NA
c-ki-ras	E	H	√	32.6	√	43.2
c-fos	E	H		71.5	√	68.1
c-raf	E	H	√	37.3	√	58.0
Histone H2A.1	E	H	√	67.4	√	94.6
α-globin	E	T	√	88.8	√	67.9
c-sis	E	T	√	77.9		NA
c-myb	E	T		45.5	√	55.6
c-fes/fps	E	T		80.3	√	NA
c-rel	E	T		26.7	√	NA
c-mos	E	T	√	74.2	√	67.3
c-fms	E	T		74.8	√	67.5
Apolipoprotein AI	E	T	√	85.0		71.3
Thy-1	E	T		79.4	√	76.4
Placental lactogen	E	T	√	74.5		43.9
Complement C4	E	T		70.8	√	65.1
Immunoglobulin Kappa constant	E	T		NA	√	59.4
Albumin	E	T		39.0	√	57.0
N-ras	E	?	√	45.7		55.7
c-abl	E	?		71.5	√	68.4
Skeletal muscle actin	L	T	√	89.1		77.1
β-globin	L	T	√	66.4	√	66.9
α1-antitrypsin	L	T	√	68.8	√	64.2
β-casein	L	T		53.0	√	49.6
Phenylalanine hydroxylase	L	T	√	52.5		56.0
Factor IX	L	T	√	35.4		31.8
Fibronectin	L	T	√	50.2		53.4
Myosin heavy α-cardiac	L	T		85.8	√	80.7
N-myc	L	T	√	79.8	√	75.3
α-amylase 1	L	T		34.1	√	40.0
Major urinary proteins	L	T		NA	√	41.3
Immunoglobulin kappa variable	L	T		56.6	√	45.2

[a]The replication time of the gene: E-early and L-late.
[b]The expression status of the gene: H-housekeeping, T-tissue specific and ?-unknown.
[c]Replication time known in primate/rodent cell line.
[d]Third postion G+C content.

with the available data for each gene. The isochore G+C contents so produced were then compared to see if any overlap existed in the range of early and late replicating fractions.

Of all the relationships between gene and isochore G+C content that have been published the best, in terms of sample size and correlation coefficient, is that given by Aissani et al (5) for human third position G+C contents. Aissani et al chose to leave out two genes from their regression analysis because one of the genes had very biased amino acid composition, and the other had a very low G+C content. Since the second of these reasons appears to be arbitary and the first is not relevent to the third position G+C content both genes were included in this study. The relationship between isochore and gene G+C content obtained by least squares linear regression is:

(1) Isochore G+C = 31.3 + 0.229×Third Position G+C

Since the error terms (residuals) appear to be normally distributed, and unrelated in magnitude or sign to the third position G+C contents, the predicted isochore G+C content for a gene of G+C content X_0 is t-distributed with N−2 degrees of freedom, a mean of Y_0, the isochore G+C content given by the regression line (1), and a standard deviation of

$$(2) \qquad S\left[1 + \frac{1}{N} + \frac{(X_0 - \bar{X})^2}{\sum (X - \bar{X})^2}\right]^{\frac{1}{2}}$$

where S is the standard deviation of the residuals and N the sample size of the data used in the regression. Thus by sampling at random from a t-distribution with the appropiate parameters it is possible to convert gene G+C contents into isochore G+C contents in a way consistent with the data of Aissani et al. The isochore G+C contents so produced can then be examined to see if early and late isochores overlap in G+C content. By repeating this procedure many times it is possible to assess how much overlap there must be between the G+C contents of early and late replicating isochores. For instance if we found that only 0.5% of a very large number of randomly produced isochore sets showed no overlap between early and late replicating fractions, then we would be able to reject the null hypothesis that all early replicating isochores are more G+C rich than late replicating isochores at the 0.5% level. This test was only applied to human genes because there are far fewer rodent genes for which isochore location is known. The test would therefore be much less powerful.

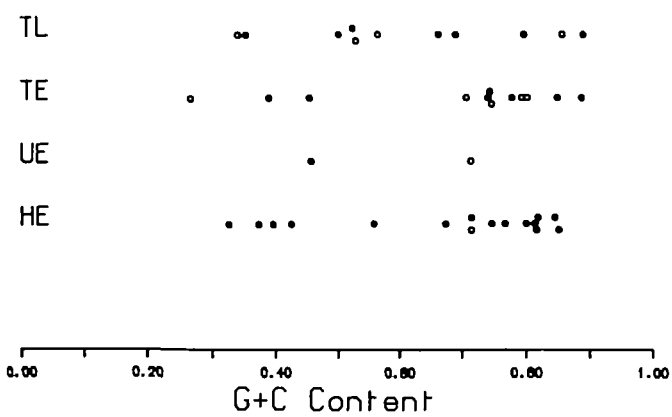

Figure 1. The third postion G+C contents of human housekeeping, tissue-specific, early and late replicating genes. HE−early replicating housekeeping genes. UE−early replicating genes of unknown expression. TE−Early replicating tissue specific genes. TL−Late replicating tissue specific genes. Filled circles are those genes whose replication time is known in a human cell line.

RESULTS

The G+C content, replication time and expression status of the 44 genes in the data set are given in table 1, and represented graphically in figures 1 and 2. Only third position G+C contents are given since the correlation between third position and isochore G+C contents is much better than for other positions (4,5). Since there is no evidence that replication times differ between rodents and primates (table 1), and no evidence of differences in the G+C contents of genes whose replication time is known and those whose replication time is inferred from another group (table 2a), it was assumed in all subsequent analyses that replication times were identical in primates and rodents. The results are not qualitatively affected by this assumption.

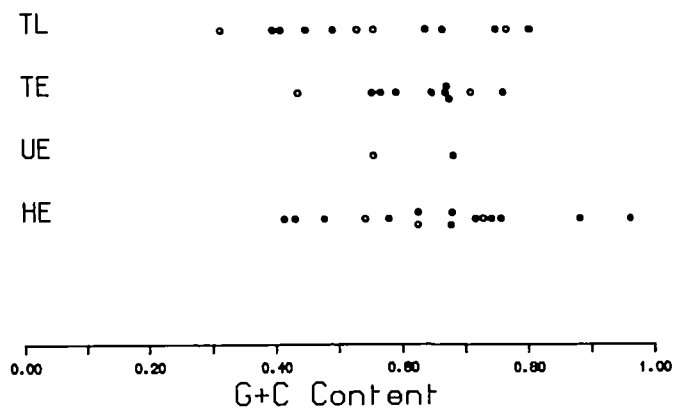

Figure 2. The third postion G+C contents of mouse housekeeping, tissue-specific, early and late replicating genes. Symbols as in figure 1. Filled circles are those genes whose replication time is known in a rodent cell line.

Table 2. Testing for differences in G+C content.

Data set	Primates	Rodents
(a)		
H	−	0.72
TE	0.14	0.90
TL	0.78	0.93
E	0.21	0.37
T	0.25	0.53
(b)		
H v TE	0.91	0.62
TE v TL	0.41	0.28
H v T	0.72	0.28
E v L	0.41	0.17

Figures in the body of the table show the probability of the two data sets being more dissimilar than they are by chance alone in a Mann-Whitney test. In part (a) the G+C contents of genes whose replication time is known in a group (e.g primates) are compared against those whose replication is inferred from another group (e.g rodents). In part (b) genes with different characteristics are compared. H−housekeeping, T−tissue specific, E−early and L−late replicating genes. The test cannot be performed for primate housekeeping genes due to insufficient sample size.

Confirming the result of Mouchiroud et al (6) figures 1 and 2 show that there is no difference in the distribution of G+C contents of housekeeping and tissue-specific genes. Mann–Whitney tests confirm this (table 2b). More importantly there is also little difference in the distributions of early and late replicating genes. The early replicating genes appear to be slightly more G+C rich than the late replicating genes but this difference is not significant (table 2b).

To illustrate how inconsistent these results are with the replication of only G+C rich isochores early, and G+C poor isochores late in S phase, isochore G+C content is plotted against third position G+C content for a set of 21 human genes (5) in figure 3. There is no horizontal line which would split the data so they look like the patterns in figures 1 and 2. For instance let us imagine that all isochores above 43% replicate early in S phase with the rest replicating late: there is almost no overlap between the G+C contents of early and late replicating genes.

It is possible to make this argument more quantitative by converting gene G+C contents to isochore G+C contents in a way consistent with the data of Aissani et al (5,figure 3), as detailed in the materials and methods. In 10000 simulated sets of isochores generated from the human early and late sets of genes, there was not a single case when the most G+C rich late replicating isochore was less G+C rich than the least G+C rich early replicating isochore; i.e there was always some overlap between the early and late replicating isochores. We can therefore reject the hypothesis that all early replicating isochores are more G+C rich than late replicating isochores at 0.05% significance or lower. Furthermore in 9999 cases the upper quartile (the value above which 25% of the observations lie) of the late replicating genes was greater than the lower quartile (the value below which 25% of the observations lie) of the early replicating genes. In other words the overlap was always substantial. When the test was repeated on just early and late replicating tissue specific genes there was always an overlap of G+C contents, and in all but 22 cases the early replicating lower quartile was less than the late replicating upper quartile.

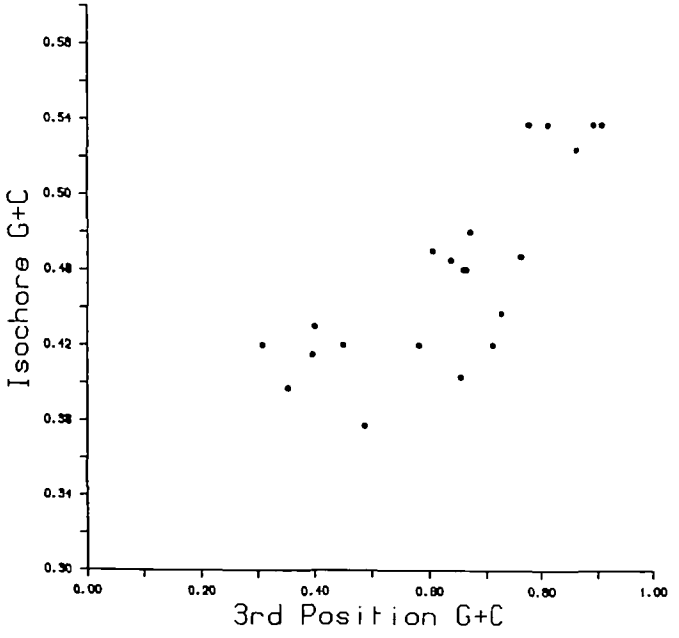

Figure 3. Isochore G+C content against third position G+C content for 21 human genes. Data from Aissani et al (5).

DISCUSSION

These results demonstrate that there is considerable heterogeneity in the G+C content of isochores replicated early and late during S phase. It is unclear however whether the replication of G+C rich and poor DNA is temporally separated, but over a much shorter time scale than the length of S phase, or whether sequences of different G+C content are simultaneously replicated. Several groups have looked at the G+C content of DNA being replicated at hourly intervals during S phase. Comings (9) found in a hamster cell line that the average G+C content of replicating DNA changed continuously from relatively A+T rich, to G+C rich before decreasing again to G+C poor. Thus there was an overlap in the G+C content of sequences replicated early and late in S phase. In contrast Tobia et al (7) and Flamm et al (8) found that in a mouse cell line the G+C content of replicating DNA decreased monotonically during S phase. However in all analyses the range of average G+C contents replicated at different times (~5%) was insufficient to cover the range of isochore G+C contents (~9–15%). It therefore seems likely that sequences of very different G+C content are replicated simultaneously during S phase.

Isochore Replication Times

Further evidence that G+C rich and poor isochores replicate in both halves of S phase is provided by a few genes for which isochore location and replication time are known (table 3). In mice there appears to be no relationship between replication time and isochore class, and in humans early replicating genes are located in all isochore classes except the very G+C poorest.

The Maintainence Of Isochores

The fact that sequences of very different G+C contents may be replicated simultaneously has some implications for the

Table 3. Gene replication time and isochore location

Gene	Isochore G+C	Replication time
Human		
Factor IX	39.7	L
β-globin	40.3	L
HPRT	41.5	E
c-mos	43.7	E
c-myc	46.7	E
Glucose-6-dehydrogenase	52.4	E
c-Ha-ras	53.7	E
α-globin	53.7	E
c-sis	53.7	E
Mouse		
IgK Variable	40.5	L
IgK Constant	42.0	E
β-globin	42.0	L
α-globin	49.1	E
Skeletal actin	49.1	L
c-abl	49.1	E

Replication time data comes from references cited in the methods section. Isochore location data comes from (5) for humans, and from (4) for mice.

maintainence of isochores. The mechanism by which isochores are maintained is the subject of considerable debate (13,22,26). The simplest and most cogent hypothesis has been put forward by Wolfe and colleagues (13,14). They proposed that different replicons are replicated in free nucleotide pools of different compositions which biases the pattern of mutation, thus producing replicons/isochores of different G+C contents. This very neatly explains the relationship between replication time and G+C content that was originally thought to exist, since it had been shown that the free nucleotide pool composition changed through the cell cycle (27,28). The fact that isochores of different G+C contents appear to replicate simultaneously poses something of a problem for this hypothesis, unless the free nucleotide pools are spatially heterogeneous. Paradoxically one is loath to drop the Wolfe/Li/Sharp hypothesis because it provides a very elegant explanation of the correlation between gene and isochore G+C contents, one of the observations which led to the original paradox. The correlation arises under this hypothesis, because although selection and DNA repair may vary across a replicon, all sequences in a replicon have the same pattern of misincorporation which is different to other replicons replicated under different conditions. Therefore sequences within a replicon are expected to have correlated compositions.

It should be appreciated that the conclusions reached via table 1 are only strictly applicable to the cell lines in which the gene replication times were studied. The conclusions do not neccessarily extend to the germ-line, which is the relevent tissue when discussing the origins and maintainence of isochores. It is possible that the pattern of replication is quite different in germ and somatic cell lines. However it is clear from the present work that in certain cell lines both G+C rich and G+C poor isochores replicate early and late in the cell cycle.

ACKNOWLEDGEMENTS

I would like to thank Paul Sharp, Peter Keightley, Bill Hill and two anonymous referees for their encouragement and comments on this manuscript. I would also like to thank the SERC for their financial assistance.

REFERENCES

1. Comings,D.E. (1978) *Ann Rev Genet.*, **12**, 25–46
2. Goldman,M.A. (1988) *Bioessays* **9**, 50–55
3. Holmquist,G.P. (1989) *J Mol Evol* **28**, 469–486
4. Bernardi,G., Olofsson,B., Filipski,J., Zerial,M., Salinas,J., Cuny,G., Meunier-Rotival,M. and Rodier,F. (1985) *Science* **228**, 953–958
5. Aissani,B., D'Onofrio,G., Mouchiroud,D., Gardiner,K., Gautier,C. and Bernardi,G. (1991) *J Mol Evol* **32**, 493–503
6. Mouchiroud,D., Fichant,G. and Bernardi,G. (1987) *J Mol Evol* **26**, 198–204
7. Tobia,A.M., Schildkraut,C.L. and Maio,J.J. (1970) *J Mol Biol* **54**, 499–515
8. Flamm,W.G., Bernheim,N.J. and Brubaker,P.E. (1971) *Exp Cell Res* **64**, 97–104
9. Comings D.E (1971) *Exp Cell Res* **71**, 106–112
10. Hutchison,H.T. and Gartler,S.M (1973) *Tex Rep Biol and Med* **31**, 321–329
11. Holmquist,G.P., Gray,M., Porter,T. and Jordan,J. (1982) *Cell* **31**, 121–129
12. Bernardi,G. (1989) *Ann Rev Genet* **23**, 637–661
13. Wolfe,K.H., Li,W-H. and Sharp,P.M. (1989) *Nature* **337**, 283–285
14. Wolfe,K.H. (1991) *J Theor Biol* **149**, 441–451
15. Goldman,M.A., Holmquist,G.P., Gray,M.C., Caston,L.A. and Nag,A. (1984) *Science* **224**, 686–692
16. Calza,R.E., Eckhardt,L.A., DelGiudice,T. and Schildkraut,C.L. (1984) *Cell* **36**, 689–696
17. Iqbal,M.A., Plumb,M., Stein,G. and Schildkraut,C.L. (1984) *Proc Natl Acad Sci USA* **81**, 7723–7727
18. Iqbal,M.A., Chinsky,J., Didamo,V. and Schildkraut,C.L. (1987) *Nucleic Acid Res* **15**, 87–103
19. Hatton,K.S., Dhar,V., Brown,E.H., Mager,D. and Schuldkraut,C.L. (1988) *Mol Cell Biol* **8**, 2149–2158
20. Devereux,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acid Res* **12**, 387–395
21. Mouchiroud,D., Gautier,C. and Bernardi,G. (1988) *J Mol Evol* **27**, 311–320
22. Bernardi,G., Mouchiroud,D., Gautier,C. and Bernardi,G. (1988) *J Mol Evol* **28**, 7–18
23. McAlpine,P.J., Stranc,C.C., Boucheix,C. and Shows,T.B. (1990) *Cytogenet Cell Genet* **55**, 5–76
24. Davisson,M.T., Lalley,P.A., Doolittle,D.P. and Hillyard,A.L. (1990) *Cytogenet Cell Genet* **55**, 434–456
25. Hillyard,A.L., Doolittle,D.P., Davisson,M.T. and Roderick,T.H. (1991) *Mouse Genome* **89**, 16–30
26. Filipski,J., Salinas,J. and Rodier,F. (1987) *DNA* **6**, 109–118
27. McCormick,P.J., Danhauser,L.L., Rustim,Y.M. and Bertram,J.S. (1983) *Biochim Biophys Acta* **755**, 36–40
28. Leeds,J.M., Slabaugh,M.B. and Matthews,C.K. (1985) *Mol Cell Biol* **5**, 3443–3450