256–260 Nucleic Acids Research, 1995, Vol. 23, No. 2

One short well conserved region of *Alu*-sequences is involved in human gene rearrangements and has homology with prokaryotic *chi*

Niels S. Rüdiger*, Niels Gregersen[1] and Morten C. Kielland-Brandt[2]

Department of Clinical Biochemistry and [1]Center for Medical Molecular Biology, Aarhus University Hospital, Skejby Hospital, DK-8200 Aarhus N, Denmark and [2] Department of Yeast Genetics, Carlsberg Laboratory, Copenhagen, Valby, Denmark

Received September 27, 1994; Revised and Accepted December 14, 1994

ABSTRACT

***Alu* elements have repeatedly been found involved in gene rearrangements in humans. Although these elements have been suggested to stimulate gene rearrangements, sparse information is available for the possible mechanism(s) of these events. Here we present a compilation of *Alu* elements that have been involved in recombinational events leading to gene rearrangements, indicating the presence of a common 26 bp core sequence at or close to the sites of recombination. Besides the obvious possibility of retrotransposition, gene rearrangements may be induced by sequences that stimulate genetic recombination. We suggest that the core sequence stimulates recombination and may thereby cause the frequent involvement of these elements in gene rearrangements. Curiously, the core sequence contains the pentanucleotide motif CCAGC, which is also part of *chi*, an 8 bp sequence known to stimulate *recBC* mediated recombination in *Escherichia coli*.**

INTRODUCTION

Disease causing gene rearrangements have been demonstrated in several genes. Although all genetic rearrangements are nonhomologous recombinations in a global sense, gene rearrangements are often classified as being due to homologous or non-homologous recombination, based upon the presence or absence, respectively, of nucleotide sequence homology between the parental sites of recombination. In the present paper we shall use the terms homologous recombination and non-homologous recombination in this subgenic sense.

The *Alu* family of short interspersed repetitive DNA elements (SINEs) is only found within the genomes of primates, which evolved within the last sixty-five million years (1). Taken together with the preponderance of *Alu* elements this indicates that *Alu* elements are mobile sequences. Moreover, the *Alu* elements found in the untranslated part of the human LDL-receptor gene are not found in the corresponding gene in lower animals, such as baboons, cattle, and rabbits, but investigations have suggested their occurrence in chimpanzee and gorilla (2). Thus *Alu* elements must have been introduced and multiplied after the evolutionary divergence of the lines for humans and baboons happening some 30 million years ago (2).

In man, the repetitive family of *Alu* retroposon elements are involved in most gene rearrangements that have occurred by homologous recombination. This includes sex chromosome exchange in XX males (3) as well as in the genes for the LDL-receptor (4,5), human ß-globin (6,7), the *c-sis* protooncogene (8), α-globin (9), and apolipoprotein B (10).

We therefore initiated a search for recombinational hotspots by looking for similarities and differences in sequences that have been involved in recombinational events. Some sequence motifs, such as $d(GT{\cdot}AC)_n$ in mammalian cells (11) and yeast (12) and a RAP1 binding site in yeast (13) have been shown to invoke a stimulatory effect on recombination (11). We started out to analyse rearrangements in a single, well-known gene, coding for the LDL-receptor. Whereas most *Alu* elements are located in intergenic DNA or in introns, the LDL-receptor gene is exceptional in that *Alu* elements, besides being present in most of the 17 introns, are found in the nontranslated part of the last exon (14). Today more than 30 deletions have been located in the LDL-receptor gene in patients with familial hypercholesterolemia (15). We have previously reported a deletion of exon 5 (FH-DK 3) in the LDL-receptor gene, which was the result of a recombination between two identical 28 bp sequences (4). Each of the sequences was part of an *Alu* element present in introns 4 and 5, respectively, of the gene. The delineation of a similar, but not identical, deletion of exon 5 (FH 626) was previously reported by Hobbs *et al.* (5). Both deletions involve the same sequence from intron 5 for crossover. Although the sequences upstream for the crossovers leading to the deletions are different in intron 4, the four sequences involved in these gene rearrangements have 26 bp in common (4), pointing to the possibility that this sequence contains a recombinogenic hotspot. This 26 bp sequence was compared with sequences involved in other characterised homologous and non-homologous recombinations in the LDL-receptor gene (Fig. 1A). Most intriguingly eight out of nine rearrangements, characterised at the sequence level in this gene, involve

* To whom correspondence should be addressed

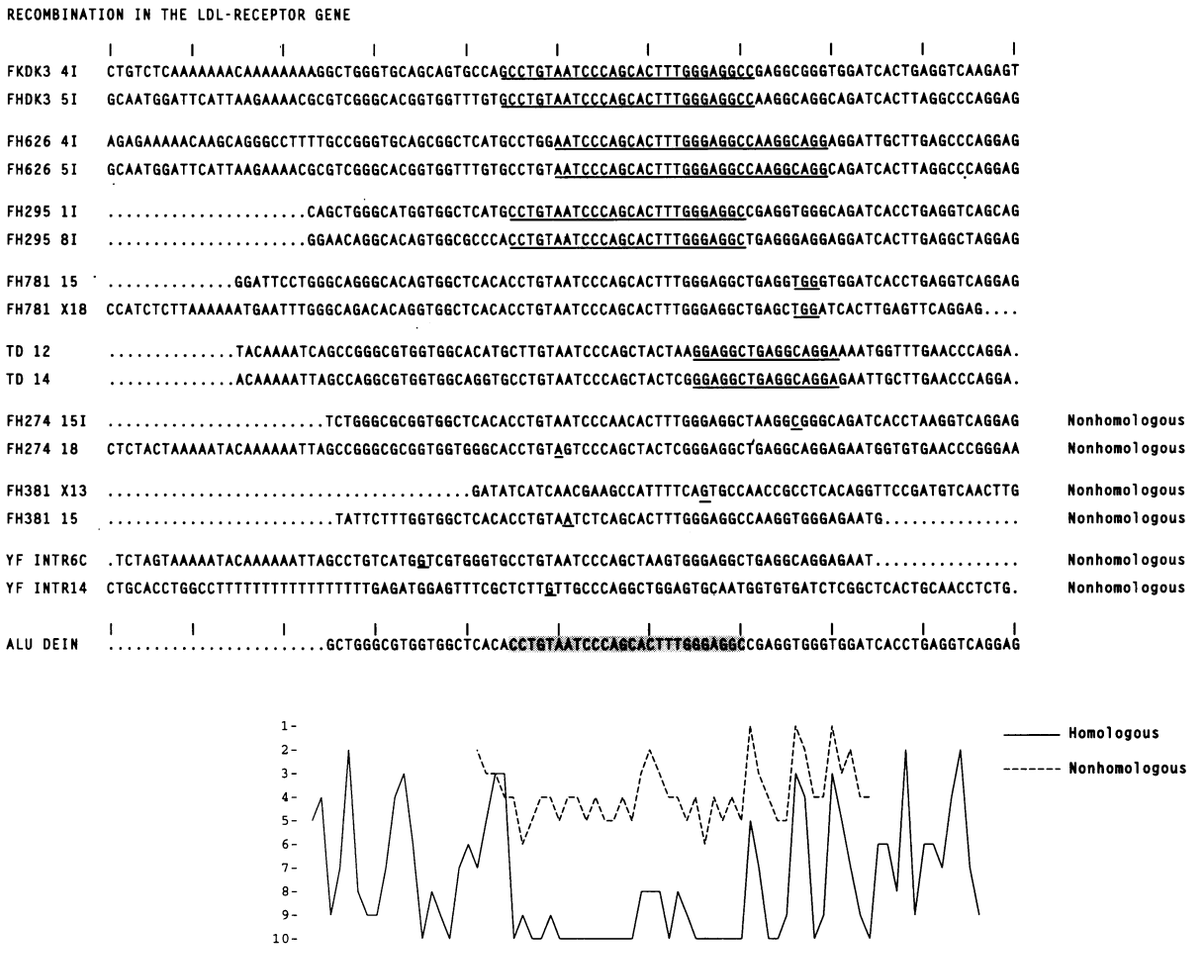

Figure 1. Gene rearrangements in the human LDL-receptor gene involving *Alu* elements. Compilation of parental sequences involved in recombinations leading gene rearrangements in the human LDL-receptor gene. Strand and orientation have been chosen for homology with the *Alu* sequence (*ALU*-DEIN) reported by Deininger *et al.* (1). Listing of parental sequences involved in eight recombinational events in the LDL-receptor gene five of which exhibit sequence homology between the parental sites of recombination while three were classified as being nonhomologous (indicated). Names of the involved strands have been denoted with a prefix referring to the patients in whom they have been found. The last number refers to the intron number in which the parental sequence was located except when an 'X' is added in which case it refers to exon number. Most sequences are represented by the upper strand in the orientation of transcription. Sequence names ending with an 'I' indicate the orientation opposite of transcription and the lower strand is presented. Below the sequences the number of identities is graphically illustrated for homologous (—) recombinations in the gene and the region with maximum identity was defined as a core sequence shaded in the *ALU* DEIN sequence. A similar graphical illustration is made for the non-homologous (--) recombinational events in the gene. In homologous recombinations the sites of crossover are within the underlined identical stretches. In the non-homologous recombinations the sites of crossover are within the underlined bases (when more bases are underlined) or just 5′ to the underlined base (when only one base is underlined). The rearrangements were the following: FH DK 3 (4), FH626 (5), FH295 (37), FH781 (38), FH274 (17) and data obtained from the EMBL data bank (AC L00352; K02573) FH381 (39), YF (40).

Alu elements. This corresponds to the involvement of 15 strands with *Alu* elements (the FH381 rearrangement involves a coding sequence) out of 18 possible for generating the nine rearrangements in the LDL-receptor gene (one rearrangement not involving *Alu* elements is not shown in Fig. 1A). Because a 26 bp core sequence is highly conserved in all 15 *Alu* elements, we suggest that this sequence is a recombinational hotspot.

RESULTS

Sequence analysis

The sequences shown in Figure 1A share one region of extensive homology, a region that includes the 26 bp core sequence. A graphical representation of the sequence identities of strands involved in the homologous (—) and non-homologous (--) recombinations, respectively, have been used to define a core sequence. The homologous recombinations (the upper 10 strands in Fig. 1) show a maximum number of identities with the shaded sequence of the *ALU*-DEIN sequence. This sequence is not conserved to quite the same degree in the six strands involved in non-homologous recombination, although the graph for the sequences involved in non-homologous recombination results in a pattern similar in shape to the graph obtained for homologous recombination. The frequent involvement of the 26 bp core sequence 5′-CCTGTAATC-CCAGCACTTTGG-GAGGC-3′ in rearrangements in the LDL-receptor gene points to a possible recombinogenic nature of the sequence. The core sequence is identical to sequences in the left arm of the consensus *Alu* element published by Deininger *et al.* (1) and homologous to a sequence

Figure 2. Gene rearrangements in man involving *Alu* elements.Compilation of parental DNA segments from other human genes involved in rearrangements that could be classified as homologous recombination. Strand and orientation have been chosen for homology with the *Alu* sequence (*ALU*-DEIN) reported by Deininger *et al.* (1). The sites of crossover are within the underlined identical stretches. When the crossover sequences are not shown the distances to the breakpoints are indicated. The rearrangements were the following: α-galactosidase A gene (GAL-) (41), apolipoprotein B (APB) (10), C_1-inhibitor (C1INHF1 C1INHF4 C1IN-A) (27,42), *c-sis* protooncogene (MEN) (8),X-chromosome (XX) (3), glycophorin A and B genes (GP) (43), β-hexosaminidase B (Sandhoff) (44) and α-globin gene (ALFA RA ME) (9).

near the middle of the element. While most of the recombining sequences in Figure 1A correspond to the former sequence, the latter sequence is equivalent to the TD-14 (16) and FH274-18 (17) sequences.

To further investigate the possible significance of the indicated core sequence, DNA sequences, which have been involved in the formation of joint fragments by homologous recombination in other human genes, were compared (Fig. 2). The facts that: (i) the presented rearrangements all have involved at least one *Alu* element (FH381 in Figure 1 involves one *Alu*-element and one coding exon sequence), and (ii) the core sequence is well conserved in the recombining strands; indicate that the participation of the core sequence in rearrangements is not limited to the LDL-receptor gene but that the core sequence is more generally involved in recombinations involving *Alu* elements.

We therefore extended the comparative studies to include rearrangements occurring by recombination between non-homologous sequences involving at least one *Alu* element (Fig. 3). There is in the literature a general inconsistency in the classification of non-homologous recombinational events. We have chosen to classify as non-homologous recombination only those events where the stretch of homology is less than 6 bp. Unfortunately only short stretches of sequences were available near the sites of crossover. Curiously, in all three cases at least one of the parental strands shows homology to the core sequence, suggesting that this sequence might also be important for recombinational events of the non-homologous type—an observation already made for similar events in the LDL-receptor gene (Fig. 1).

DISCUSSION

An alternative to stating a recombinogenic nature of the core sequence has been considered. It is tempting to speculate that the occurrence of these or similar sequences near the crossover sites is just a consequence of the enormous abundance of *Alu* elements in the human genome, where these elements have been estimated to occur in about 500 000 copies per haploid human genome (2), rather than a recombinogenic property of the repeats. While the abundance is likely to play a role, we favour the idea that the core sequence, or part of it, is recombinogenic. The presence of recombinogenic sequences in *Alu* elements has been suggested by others (18,19) and may well be responsible for the occurrence of free *Alu*-containing circles (20,21). *Alu* elements represent a highly diverged family (22), but a region identical to the core sequence defined in the present study was found by Kariya *et al.* (23) to be the most conserved region of the elements. In consistency with these observations we found that in 161 non-selected *Alu* elements from the GenBank database the core sequence has an average intersequence homology of 96.7%. Out of the 161 elements, 72 have sequences identical to the core sequence. As also shown by

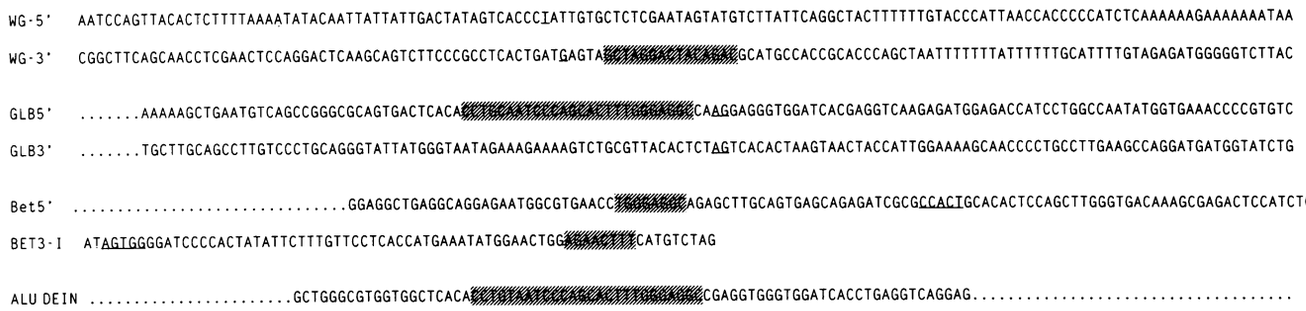

Figure 3. Human joint genes generated by non-homologous recombinations. Strand and orientation have been chosen for homology with the *Alu* sequence (*ALU*-DEIN) reported by Deininger *et al.* (1). The sites of crossover are within the underlined identical stretches. The rearrangements were the following β-hexosaminidase B (WG) (45), and β-globin gene (GLB BET 5) (6,7).

Kariya *et al.* (23) the homology outside the core sequence is considerably lower (~80%). Therefore, unless the analyses are biased by the fact that the GenBank database in particular includes sequences from genes, the core sequence is well conserved in *Alu* elements in general. This high conservation may of course explain a high incidence of homologous recombination involving the core, but without a more specialised recombinogenic function it would be hard to understand the relatively frequent occurrence of non-homologous events at or near the core.

In summary, we suggest that the core sequence, or part of it, stimulates homologous and non-homologous recombination within the core or at nearby sites. In *Escherichia coli*, the *recBC* dependent recombination system is stimulated near the so called *chi* recombinational hotspots [see review by Stahl (24)], which consist of the sequence 5′-GCTGGTGG-3′ plus its complement (25). Curiously, a motif of five consecutive base pairs of this element (5′-CCAGC-3′) is present in the *Alu* core sequence. Perhaps this structural similarity corresponds to a functional similarity (20). Four bases (CCAG) of this pentanucleotide motif have been shown by Chou and Morrison (26) to be present near non-homologous recombination breakpoints at the Ig loci.

Recombinations involving *Alu* elements may therefore be divided into two categories of events. One is retrotransposition (27,28), which results in multiplication of the elements in the genome in a manner seen for viruses, and the other in recombinations leading to exchange of genetic material between regions with homologous or sometimes non-homologous sequences. Multiplication of genomic material is unlikely to involve homologous recombination, since such a mechanism would limit the spreading of the *Alu* elements to locations in the genome that already harbor it. With few exceptions all known SINEs appear to be retroposons that usually make target site duplication of 7–21 bp upon insertion (29). These insertions differ fundamentally from the recombinations listed in Figures 1 and 2 that all have involved G/C-rich domains. The observations therefore support the notion of different mechanisms for insertion and recombination involving *Alu* elements.

Because the rearrangements in Figures 1 and 2 all possess the core sequence, we suggest that the recombination is stimulated by this sequence. If so, the recombinations of the non-homologous type in Figure 1 have been generated by a recombination machinery with preferences for certain sequences. Several indications may support such a consideration. First, if no recognition sequence preferences exist, the likelihood of non-homologous recombination between an *Alu* sequence with non-*Alu* sequences

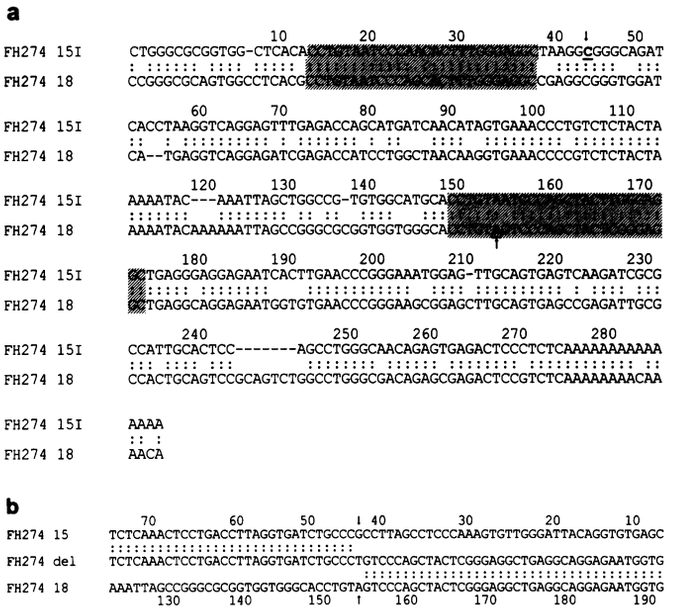

Figure 4. Analysis of a non-homologous recombination involving two *Alu* elements. Based on the available sequence data for the recombining helices in patient FH274 (17) the result of an analysis for identities is shown. (**a**) The analysis reveals a 80.3% identity in the 304 bp sequence region of the parental sequences involved in the recombination. The two sequences are numbered according to the sequence data listed by Lehrman *et al.* (17) (FH274 15I - inverted) and data obtained from the EMBL-sequence bank (FH274 18 = AC L00352; K02573). The crossover has occurred at positions 3′ of the underlined bases which are included in the rearranged gene. Sequences with homology to the core sequence in Figure 1 are shaded. (**b**) Alignment of the parental sequences around the position of crossover. The crossover is indicated with arrows. The sequences shown are from Lehrman *et al.* (17).

is about 94% [*Alu* elements constitute about 6% of the human genome (2)], and it is therefore unlikely that two out of three rearrangements of the non-homologous type involve two of these elements and one involves a single *Alu* element. Secondly, in FH 274 (17), the rearrangement has involved an *Alu* element from intron 15 and its most homologous counterpart from exon 18 (Fig. 4a). Due to the bipartite structure of *Alu* elements two sequences in the element show homology to the core sequence (shaded in Fig. 4a). As shown in Figure 4 the crossovers (FH274 15I position 43, FH274 18 position 154) have occurred between sequences harboring the core sequence. Furthermore, in the FH381 rearrange-

ment a duplication of the sequence AGCACTTTGGG, which is identical to a subset of the core sequence, has occurred 3 bp from the site of crossover generating the FH381 allele sequence TTTTC AGTCTC|AGCACTTTGGG|AGCACTTTGGG|AGGCC.

The molecular mechanism of the proposed stimulation of recombination can at present only be subject to speculation. Slipped mispairing during replication (30) is suggested to account for excision in yeast of bacterial Tn5 (almost 9 kb) in a mechanism involving a long (1.5 kb) inverted repeat (IS50) and a short (9 bp) direct repeat (31), an event reminiscent of that presented in Figure 4. In a very different class of mechanisms it has been suggested that topoisomerase II can mediate illegitimate recombination *in vitro* by double-strand cleavage and subunit exchange (32) and has also been implicated in similar events *in vivo* (33). Also, topoisomerase I has been suggested to mediate illegitimate recombination (34,35), and a specific mechanistic model has been proposed (36). It is therefore noteworthy that the *Alu* core sequence contains a putative topoisomerase I substrate site (CTT).

In conclusion the presented compilation of rearrangements supports the idea that the 26 bp core sequence in *Alu* elements functions as a recombinational hotspots. The compilation points to the possibility that there may be a common mechanism for the rearrangements classified to have taken place by homologous and non-homologous recombination involving *Alu* elements.

ACKNOWLEDGEMENTS

This work was supported by the Bioregulatory Research Center, University of Aarhus, Danish Heart Foundation and the NOVO FONDs komité.

REFERENCES

1 Deininger, P.L., Jolly, D.J., Rubin, C.M., Friedmann, T. and Schmid, C.W. (1981) *J.Mol.Biol.* **151**,17–33.
2 Hobbs, H.H., Lehrman, M.A., Yamamoto, T. and Russell, D.W. (1985) *Proc.Natl.Acad.Sci. USA* **82**,7651–7655.
3 Rouyer, F., Simmler, M-C., Page, D.C. and Weissenbach, J. (1987) *Cell* **51**,417–425.
4 Rüdiger, N.S., Hansen, P.S., Jørgensen, M., Faergeman, O., Bolund, L. and Gregersen, N. (1991) *Eur.J.Biochem.* **198**,107–111.
5 Hobbs, H.H., Brown, M.S., Goldstein, J.L. and Russell, D.W. (1986) *J.Biol.Chem.* **261**,13114–13120.
6 Henthorn, P.S., Mager, D.L., Huisman, T.H.J. and Smithies, O. (1986) *Proc.Natl.Acad.Sci. USA* **83**,5194–5198.
7 Vanin, E.F., Henthorn, P.S., Kioussis, D., Grosveld, F. and Smithies, O. (1983) *Cell* **35**,701–709.
8 Smidt, M., Kirsch, I. and Ratner, L. (1990) *J.Clin.Invest.* **86**,1151–1157.
9 Nicholls, R.D., Fischel-Ghodsian, N. and Higgs, D.R. (1987) *Cell* **49**,369–378.
10 Huang, L-S., Ripps, M.E., Korman, S.H., Deckelbaum, R.J. and Breslow, J.L. (1989) *J.Biol.Chem.* **264**,11394–11400.
11 Bullock, P., Miller, J. and Botchan, M. (1986) *Mol.Cell.Biol.* **6**,3948–3953.
12 Treco, D. and Arnheim, N. (1986) *Mol.Cell.Biol.* **6**,3934–3947.
13 White, M.A., Wierdl, M., Detloff, P. and Petes, T.D. (1991) *Proc.Natl.Acad.Sci. USA* **88**,9755–9759.
14 Südhof, T.C., Goldstein, J.L., Brown, M.S. and Russell, D.W. (1985) *Science* **228**,815–822.
15 Rüdiger, N.S., Heinsvig, E.M., Hansen, F.A., Færgeman, O., Bolund, L. and and Gregersen, N. (1991) *Clin.Genet.* **39**,451–462.
16 Horsthemke, B., Beisiegel, U., Dunning, A., Havinga, J.R., Williamson, R. and Humphries, S. (1987) *Eur.J.Biochem.* **164**,77–81.
17 Lehrman, M.A., Schneider, W.J., Südhof, T.C., Brown, M.S., Goldstein, J.L. and Russell, D.W. (1985) *Science* **227**,140–146.
18 Barsh, G.S., Seeburg, P.H. and Gelinas, R.E. (1983) *Nucleic Acids Res.* **11**,3939–3958.
19 Calabretta, B., Robberson, D.L., Barrera-Saldana, H.A., Lambrou, T.P. and Saunders, G.F. (1982) *Nature* **296**,219–225.
20 Haynes, S.R., Toomey, T.P., Leinwand, L. and Jelinek, W.R. (1981) *Mol.Cell.Biol.* **1**,573–583.
21 Rogers, J.H. (1985) *Int.Rev.Cytol.* **93**,231–271.
22 Orita, M., Sekiya, T. and Hayashi, K. (1990) *Genomics* **8**,271–278.
23 Kariya, Y., Kato, K., Hayashizaki, Y., Himeno, S., Tarui, S. and Matsubara, K. (1987) *Gene* **53**,1–10.
24 Stahl.F.W. (1979) *Ann.Rev.Genet.* **13**,7–24.
25 Smith, G.R., Kunes, S.M., Schultz, D.W., Taylor, A. and Triman, K.L. (1981) *Cell* **24**,429–436.
26 Chou, C.L. and Morrison, S.L. (1993) *J.Immunol.* **150**,5350–5360.
27 Stoppa-Lyonnet, D., Carter, P.E., Meo, T. and Tosi, M. (1990) *Proc.Natl.Acad.Sci. USA* **87**,1551–1555.
28 Wallace, M.R., Andersen, L.B., Saulino, A.M., Gregory, P.E., Glover, T.W. and Collins, F.S. (1991) *Nature* **353**,864–866.
29 Weiner, A.M., Deininger, P.L. and Efstratiadis, A. (1986) *Ann.Rev. Biochem.* **55**,631–661.
30 Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Barelle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980) *Cell* **21**,653–668.
31 Gordenin, D.A., Malkova, A.L., Peterzen, A., Kulikov, V.N., Pavlov, Y.I., Perkins, E. and Resnick, M.A. (1992) *Proc.Natl.Acad.Sci. USA* **89**,3785–3789.
32 Bae, Y-S., Kawasaki, I., Ikeda, H. and Liu, L.F. (1988) *Proc.Natl.Acad.Sci. USA.* **85**,2076–2080.
33 Sperry, A.O., Blasquez, V.C. and Garrard, W.T. (1989) *Proc.Natl.Acad.Sci. USA* **86**,5497–5501.
34 Konopka, A.K. (1988) *Nucleic Acids Res.* **16**,1739–1758.
35 Schiestl, R.H., Dominska, M. and Petes, T.D. (1993) *Mol.Cell.Biol.* **13**,2697–2705.
36 Bullock, P., Champoux, J.J. and Botchan, M. (1985) *Science* **230**,954–958.
37 Lehrman, M.A., Goldstein, J.L., Russell, D.W. and Brown, M.S. (1987) *Cell* **48**,827–835.
38 Lehrman, M.A., Russell, D.W., Goldstein, J.L. and Brown, M.S. (1987) *J.Biol.Chem.* **262**,3354–3361.
39 Lehrman, M.A., Russell, D.W., Goldstein, J.L. and Brown, M.S. (1986) *Proc.Natl.Acad.Sci. USA* **83**,3679–3683.
40 Fuscoe, J.C., Zimmerman, L.J., Lippert, M.J., Nicklas, J.A., O'Neill, J.P. and Albertini, R.J. (1991) *Cancer Res.* **51**,6001–6005.
41 Kornreich, R., Bishop, D.F. and Desnick, R.J. (1990) *J.Biol.Chem.* **265**,9319–9326.
42 Ariga, T., Carter, P.E. and Davis, A.E.III (1990) *Genomics* **8**,607–613.
43 Kudo, S. and Fukuda, M. (1989) *Proc.Natl.Acad.Sci. USA* **86**,4619–4623.
44 Neote, K., McInnes, B., Mahuran, D.J. and Gravel, R.A. (1990) *J.Clin.Invest.* **86**,1524–1531.
45 Myerowitz, R. and Hogikyan, N.D. (1987) *J.Biol.Chem.* **262**,15396–15399.