The EMBO Journal vol.3 no.2 pp.383–387, 1984

Nucleotide sequence of the human c-*myc* locus: provocative open reading frame within the first exon

Claude Gazin, Stéphane Dupont de Dinechin, Annie Hampe, Jean-Michel Masson, Patrick Martin[1], Dominique Stehelin[1] and Francis Galibert*

Laboratoire d'Hématologie Expérimentale, Centre Hayem, Hôpital Saint-Louis, 2 place du Dr. Fournier, 75475 Paris Cédex 10, and [1]Laboratoire d'Oncologie Moléculaire, Institut Pasteur, 15 rue C. Guérin, 59019 Lille Cédex, France

*To whom reprint requests should be sent

Communicated by A.van der Eb

The nucleotide sequence of a *Hind*III-*Eco*RI DNA fragment, 8 kbp long, of a λ recombinant containing the whole human c-*myc* gene has been deduced by the method of Maxam and Gilbert. This fragment encodes the complex c-*myc* locus and the sequence provides information relative to the 2.7 kb long c-*myc* transcript. It appears that although exons 2 and 3 would code for a 48-K protein homologous to the *myc* domain of the viral p110 *gag-myc* protein, the first exon, which has a large open reading frame ending with a stop codon just upstream from the donor splice site, could code on its own for a 20-K protein. Speculations about the role of that putative protein on the regulation of the expression of exons 2 and 3 are made.

Key words: cloning/human c-*myc* gene/sequencing

Introduction

During the past few months, the cellular *onc* sequences homologous to the viral *onc* genes have been intensively studied. In some cases, point mutation (Tabin *et al.*, 1982; Reddy *et al.*, 1982; Taparowsky *et al.*, 1982; Capon *et al.*, 1983) or gene rearrangement (Rechavi *et al.*, 1982; De Klein *et al.*, 1982; Dalla-Favera *et al.*, 1983) in the c-*onc* sequences have been detected and these could be responsible for, or associated with, their role in cellular malignancy. It was also shown that the NIARD sequence, which is located close to the immunoglobulin gene after chromosome 12/15 translocation leading to murine plasmocytoma, was indeed the c-*myc* gene and that the translocation break-point was within or close to the first known exon (Harris *et al.*, 1982; Stanton *et al.*, 1983; Shen-Ong *et al.*, 1982). In the case of Burkitt's lymphoma, it was shown by restriction mapping that the human c-*myc* translocates within the immunoglobulin heavy chain locus (Taub *et al.*, 1982; Adams *et al.*, 1983; Hamlyn and Rabbits, 1983). To further our understanding of *myc* gene expression in normal and pathological situations, we undertook the nucleotide sequence analysis of a λ recombinant bearing the whole human c-*myc* gene. While our work was in progress, several reports of parts of this sequence have been published, as well as complementary information on the mRNA, allowing a better interpretation at the nucleotide level (Colby *et al.*, 1983; Stanton *et al.*, 1983; Watt *et al.*, 1983).

The main additional information derived from the sequence presented here relates to the identification of the first exon and to its putative coding capacity. By itself, the first exon could code for a protein of 20 K whose expression would limit that of the 48-K protein coded by exons 2 and 3. An alternative model of expression by readthrough or use of different splice sites could allow the synthesis of a protein of 68 K with putative transforming activity.

Results

A restriction map of the 19-kb insert was established (Figure 1a) which confirms the presence of an active c-*myc* gene rather than a pseudogene. The nucleotide sequence of an 8-kbp *Hind*III-*Eco*RI DNA fragment containing the human c-*myc* gene was derived by the chemical method of Maxam and Gilbert (1980). The strategy of sequencing and the sequence are shown in Figures 1b and 2, respectively. Figure 3 shows the position of the various stop and ATG triplets along the sequence in the three reading frames of both DNA strands.

Discussion

Comparison of the sequence presented in Figure 2 with the mouse and human cDNA *myc* sequences (Stanton *et al.*, 1983; Watt *et al.*, 1983) allows us to define the position of three exons: exon 1 from position N_1 to 2881; exon 2 from position 4506 to 5277 and exon 3 from position 6654 to N_2. As shown in Figure 3 the three exons coincide with open reading frames. The 5′ position of exon 1 is difficult to assess. The cDNA sequence published by Watt *et al.* (1983) starts at position 2342. Since their sequence was derived from a cloned DNA recombinant obtained by reverse transcription and S1 digestion of a human c-*myc* mRNA, the mRNA sequence must extend upstream from that position. In that respect, it is interesting to note the presence at position 2048 of a sequence GAGCAGCA, complementary to the 5′ end of the published cDNA sequence (Watt *et al.*, 1983). This could therefore represent the 5′ end of the single-stranded cDNA copy, the short complementary sequence having served as primer for the synthesis of the second strand. This would indicate a minimum length for the first exon. Upstream sequences resembling transcriptional initiation signals are found at positions 1811 (CAT box) and 1836 (TATA box) suggesting a cap site at position 1857 or 1869. The 3′ limit of exon 3 is also difficult to predict since two potential polyadenylation site signals exist at positions 7511 and 7652. However, given these limits of error, the three exons would generate a transcript of 2.7–2.8 kb, in agreement with or slightly in excess of published values (Erikson *et al.*, 1983; Maguire *et al.*, 1983; Adams *et al.*, 1983), suggesting that no other exon exists.

Nucleotide sequence comparison between the human and mouse first exons (Stanton *et al.*, 1983) shows 237 identical nucleotides out of 434 scored positions, with two regions of homology better than 80% (between 2481 and 2535; 2685 and 2730). Comparison of the second exon shows a homology of 78%. Data available on the murine third exon are limited to

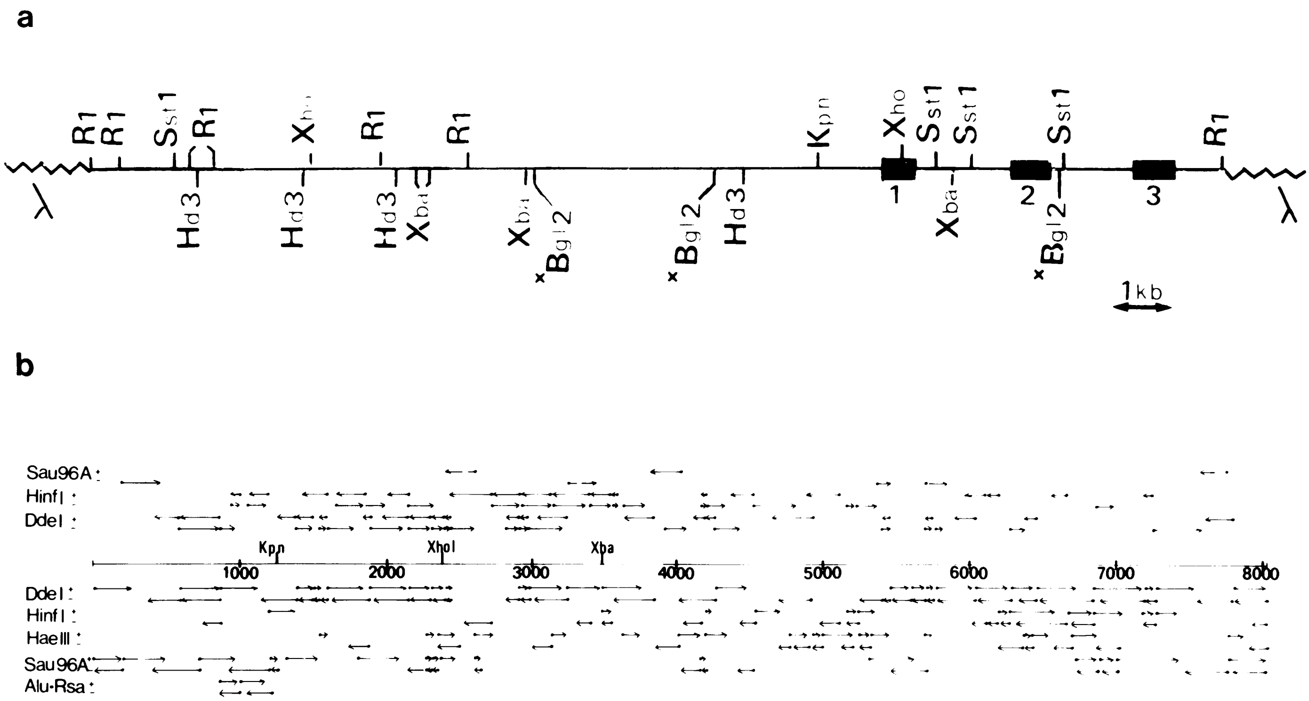

Fig. 1. (**a**) Restriction analysis of the 19-kb insert confirming the presence of an active c-*myc* gene; (**b**) diagram of analyzed DNA fragment. Length of the arrows is relative to the number of analyzed nucleotides: + indicates the strand corresponding to the polarity of the mRNA and – corresponds to the antisense strand. Fragments in the upper part of the figure were 3′ labelled, underneath fragments were 5′ labelled.

only 45 nucleotides at the 5′ extremity. Although they do not allow extensive comparison, these data seem to indicate that the degree of conservation for the third exon is about the same as that for the first exon. Comparison between the human and murine first introns indicates a degree of homology of ~40% with some small regions being better conserved, among which are the 5′ and 3′ boundaries. Comparison of the human c-*myc* sequence of Figure 3 with the previously published partial human sequences shows few differences. The sequence published by Watt *et al.* (1983) contains three additional nucleotides at positions 2435, 2642 and 2652 which close all reading frames within the first exon. As shown in Figure 4 these nucleotides are clearly missing in our sequence. These three additional nucleotides could: (a) reflect some polymorphism of the 1st exon; (b) correspond to specific insertions related to the leukemic status of the K562 cells from which Watt *et al.* (1983) obtained this cDNA; (c) correspond to some errors due either to cloning or sequencing.

According to our sequence data, the first exon has an open reading frame (ORF) starting at position 2109. In phase with that ORF there is an ATG at position 2304, which is preceded at position –3 by a purine and could be used as a start for translation (Kozak, 1981). This ORF is closed at position 2868 with a TAG stop codon located just before the donor splice site at position 2881. The ORF could, therefore, code by itself for a protein. The first question raised by this ORF concerns its actual coding capacity. When codon usage in the ORF is compared with that of the two other frames, it is apparent that some preferential choices are made. Using the parameters given by Fickett (1982), the ORF has a high probability of being a coding sequence. If this first exon indeed codes for a protein, which is also suggested by the size of the ORF and the position of the ATG triplet, the predicted protein will be 188 amino acids long with a theoretical mol. wt. of 20 931 daltons. However, it is clear that the expression of a 20-K protein by the 1st exon has to be directly demonstrated. Since there are arguments in favor of that hypothesis it is important to consider its implications.

Because of the polycistronic nature of the messenger as it appears from sequence study, the expression of the 20-K protein would prevent the expression of the 48-K protein encoded by exons 2 and 3 unless, as for the *gag-pol* protein of retroviruses (Philipson *et al.*, 1978; Murphy *et al.*, 1978; Schwartz *et al.*, 1983), a readthrough occurs, passing stop codon TAG 2868 and allowing a limited expression of these two exons.

Another hypothesis for the expression of exon 2 and 3 can be proposed. Just in front of stop codon TAG 2868, there is a putative donor splice sequence (Mount, 1982) limited by G 2866 while an acceptor consensus sequence (Mount, 1982) can be observed at position 4550 (44 nucleotides downstream from the acceptor site of exon 2, as determined by sequencing the cDNA cloned from K562 cells mRNA) (Watt *et al.*, 1983). Splicing the mRNA at these two positions would link exons 1 and 2 in such a way that translation would continue from exon 1 to exons 2 and 3 allowing the synthesis of a 68-K protein.

The amino acid sequences of the 20-K and 48-K proteins have been checked for homology with the sequences contained in the NBRF data bank. Two stretches of 80 amino acids for the 20-K protein and one for the 48-K protein were found to have a homology of 21% with a segment of the E1b protein of adenovirus 7 and E1a protein of adenovirus 12, respectively. One could speculate by analogy that proteins 20 K and 48 K have a synergistic effect upon transformation, but for the present that remains pure speculation. The percentage of homology is rather low and other proteins, some of them unrelated to the problem of transformation, were identified during this comparison (human α1 collagene III for the 20-K and Sarc RSV, bovine α S2 casein, mouse precursor α fetoprotein for the 48-K).

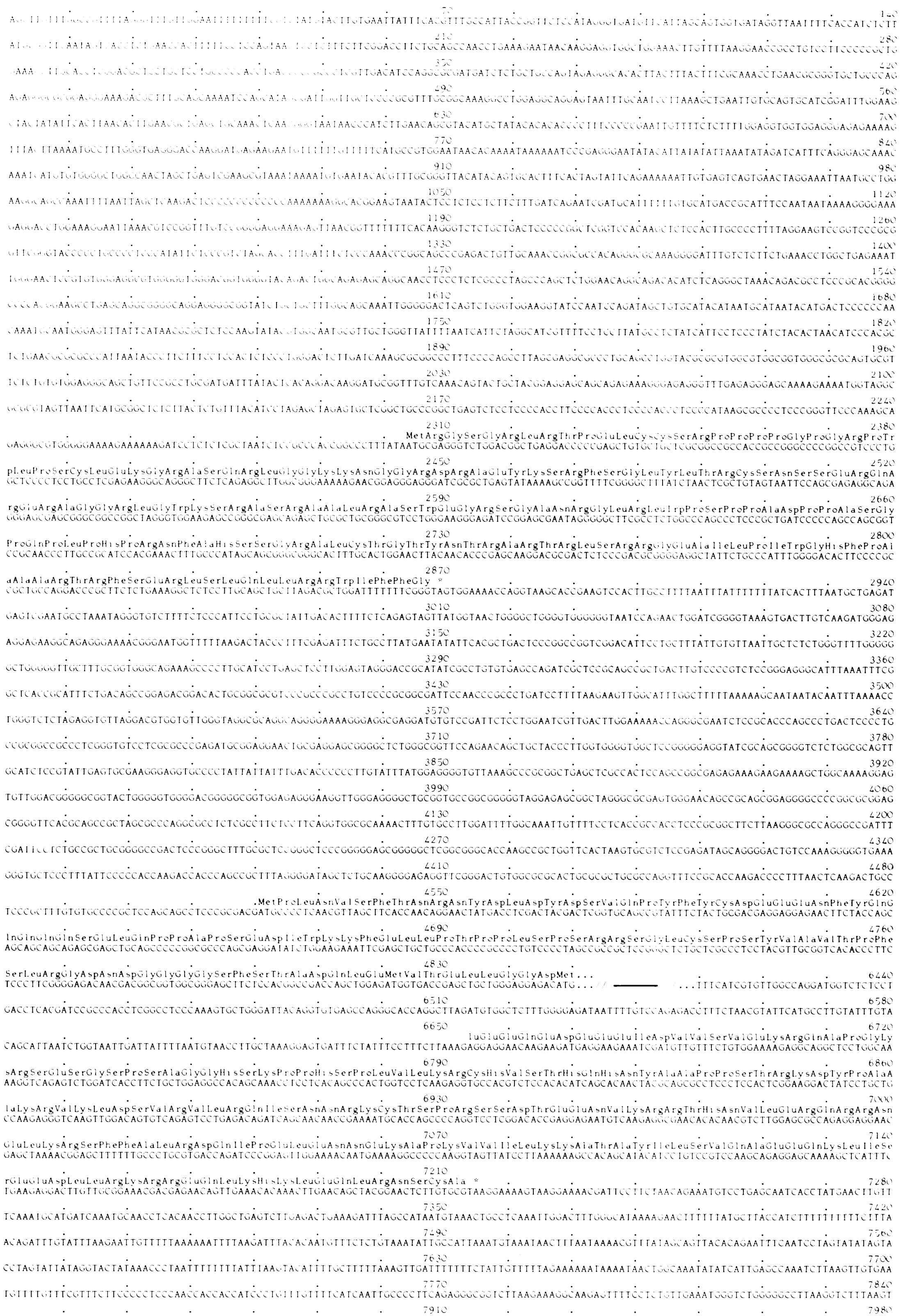

Fig. 2. Nucleotide sequence of the *Hind*III-*Eco*RI fragment (8082 bp long) corresponding to the human c-*myc* gene. The segment shown has the polarity of the c-*myc* mRNA. The sequence previously published by Colby *et al.* (1983) goes from nucleotides 3507 up to 7559. The two sequences are identical. Therefore, to gain space, the sequence between nucleotides 4850 and 6410 corresponding to nucleotides 1345 and 2905 in the sequence published by Colby *et al.* (1983) has been deleted. The amino acid sequences of the three identified exons are indicated.

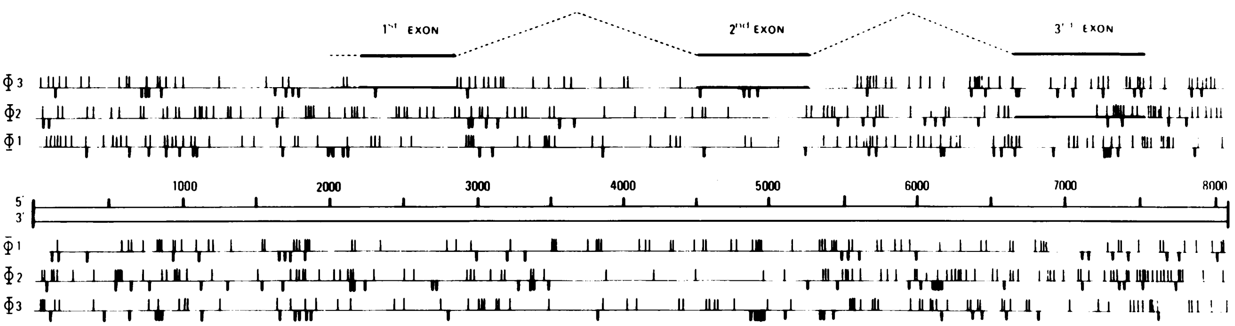

Fig. 3. Diagram showing the position of the various ORFs. Horizontal lines represent the six possible reading frames of the two DNA strands. Upper vertical bars represent stop codons. Lower vertical bars are for ATG triplets. As can be observed most of the minus strand is blocked.

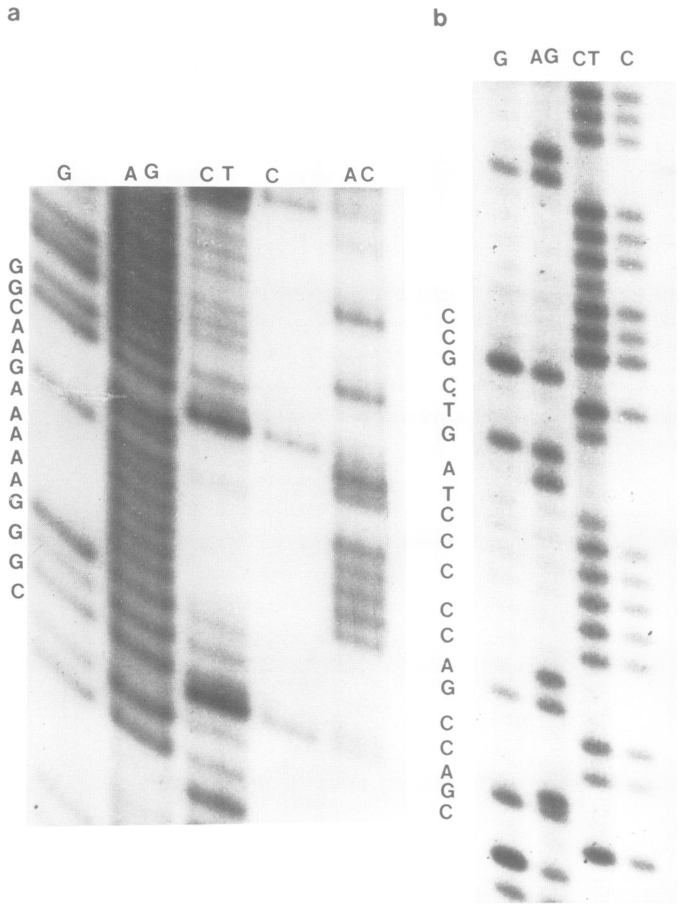

Fig. 4. Autoradiograms of 8% Maxam and Gilbert sequencing gels showing positions within the first exon where the presented sequence diverges from the sequence published by Watt *et al.* (1983): (a) part of the 5′-labelled *Hae*III fragment (2370–2535) spanning from positions 2423 to 2444. The region of discrepancy is indicated, the corresponding sequence published by Watt *et al.* being: 5′ CGGGAAAAAAGAACGG 3′; (b) part of the 3′-labelled *Hinf*I fragment (2172–2749) spanning from positions 2628 to 2659. The region of discrepancy is indicated, the corresponding sequence, published by Watt is: 3′ CGACCGGACCCCCTAGTTCGCC 5′.

Taking into account the puzzling structure of the human c-*myc* mRNA as deduced from the gene sequence, we would like to propose a working hypothesis for the expression of this oncogene. In differentiated tissues, the *myc* mRNA would essentially direct the synthesis of a 20-K protein with no transforming activity by itself. Minor products could be a 68-K protein obtained either by readthrough or alternative splicing, and a 48-K protein obtained by occasional initiation at the first ATG codon of exon 2. Such products might reach physiologically active levels in some particular stages of development (e.g., hematopoiesis) through increased transcription or regulation at the splicing level (Westin *et al.*, 1982; Coll *et al.*, 1984). According to this hypothesis, malignancy could be due to the increased expression of exons 2 and 3 whose product is homologous to the 48-K polypeptide domain coded by transforming avian retroviruses transducing exons 2 and 3 (Alitalo *et al.*, 1983; Watson *et al.*, 1983). Such an increased expression of exons 2 and 3 might be attained through abnormalities in the regulation of the mechanisms described above, or by translocation resulting in damage to the first exon and causing translation to start with ATG 4521 located at the 5′ end of exon 2.

Materials and methods

Selection of human c-myc recombinant

Using a radioactive v-*myc* RNA probe, a human DNA library obtained after partial *Eco*RI digestion (Lawn *et al.*, 1978) was screened (Leprince *et al.*, 1983). Plaques giving a positive signal were purified and amplified. One of them containing a 18-kbp fragment was selected. *Hind*III-*Eco*RI fragment hybridizing to the viral probe was subcloned in plasmid pUC9. This recombinant was used as starting material for the sequencing study.

Nucleotide sequencing

Nucleotide sequencing of pUC9-*myc* clone was performed by the method of Maxam and Gilbert (1980) using five different chemical reactions specific for G, AG, CT, C and AC. Usually ~10 pmol DNA was restricted with a given endonuclease, and the fragments were dephosphorylated and labelled with [γ-^{32}P]ATP and polynucleotide kinase as described (Hérissé *et al.*, 1980). To separate the two labelled ends, fragments were denatured at 92°C in 30% dimethyl sulfoxide and fractionated in polyacrylamide (Maxam and Gilbert, 1980). Fragments with recessed 3′ ends were labelled with a [α-^{32}P]NTP by use of DNA polymerase I (Hartley and Donelson, 1980). Both strands were independently analyzed and in several regions on both strands in two directions. All restriction sites used as starting points were also analyzed as internal sites within overlapping fragments, to detect very small fragments produced by potentially proximal cleavage events.

Acknowledgements

We thank Dr. Tom Maniatis for the gift of the human DNA library, Dr. Alan Kay and Luc d'Auriol for their critical reading of the manuscript, and Dr. Alan Kay for his help with the computer. This work was supported by the 'Institut National de la Santé et de la Recherche Médicale (INSERM)' and the 'Fondation contre la Leucémie'.

References

Adams,J.M., Gerondakis,S., Webb,E., Corcoran,L.M. and Cory,S. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 1982-1986.

Alitalo,K., Bishop,J.M., Smith,D.H, Chen,E.Y., Colby,W.W. and Levinson,A.D. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 100-104.

Capon,D.J., Chen,E.Y., Levinson,A.D., Seeburg,P.H. and Goeddel,D.V. (1983) *Nature*, **302**, 33-37.

Colby,W.W., Chen,E.Y., Smith,D.H. and Levinson,A.D. (1983) *Nature*, **301**, 722-725.

Coll,J., Saule,S., Martin,P., Raes,M.B., Lagrou,C., Gras,T., Beug,H., Simon,I.E. and Sthéhélin,D. (1984) *Exp. Cell Res.*, in press.

Dalla-Favera,B., Martinotti,S., Gallo,R.C., Erickson,J. and Croce,C.M. (1983) *Science (Wash.)*, **219**, 963-967.

De Klein,A., Van Kessel,A.G., Grosveld,G., Bartram,C.R., Hagemeijer,A., Bootsma,D., Spurr,N.K., Heisterkamp,N., Groffen,J. and Stephenson, J.R. (1982) *Nature*, **300**, 765-767.

Erikson,J., Ar-Rushdi,A., Drwinga,H.L., Nowell,P.C. and Croce,C.M. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 820-824.

Fickett,J.W. (1982) *Nucleic Acids Res.*, **10**, 5303-5318.

Hamlyn,P.H. and Rabbits,T.H. (1983) *Nature*, **304**, 135-139.

Harris,L.J., Lang,R.B., Marcu,K.B. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 4175-4179.

Hartley,J.L. and Donelson,J.E. (1980) *Nature*, **286**, 860-864.

Hérissé,J., Courtois,G. and Galibert,F. (1980) *Nucleic Acids Res.*, **8**, 2173-2191.

Kozak,M. (1981) *Nucleic Acids Res.*, **10**, 5303-5252.

Lawn,R.M., Fritsch,E.F., Parker,R.C., Blake,G. and Maniatis,T. (1978) *Cell*, **15**, 1157-1174.

Leprince,D., Saule,S., De Taisne,C., Gegonne,A., Begue,A., Righi,M. and Stéhélin,D. (1983) *EMBO J.*, **2**, 1073-1078.

Maguire,R.T., Robins,T.S., Thorgeirsson,S.S. and Heilman,C.A. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 1947-1950.

Maxam,A. and Gilbert,W. (1980) *Methods Enzymol.*, **65**, 499-460.

Mount,S.M. (1982) *Nucleic Acids Res.*, **10**, 459-472.

Murphy,E.C., Kopchickj,J., Watson,K.F. and Arlinghous,R.B. (1978) *Cell*, **13**, 359-369.

Philipson,L., Andersson,P., Olschevsky,U., Weinberg,R., Baltimore,D. and Gesteland,R. (1978) *Cell*, **13**, 189-199.

Rechavi,G., Givol,D. and Canaani,E. (1982) *Nature*, **300**, 607-610.

Reddy,E.P., Reynolds,R., Santos,E. and Barbacid,M. (1982) *Nature*, **300**, 149-152.

Schwartz,D.E., Tizard,R. and Gilbert,W. (1983) *Cell*, **32**, 853-869.

Shen-Ong,G.L.C., Keath,E.J., Piccoli,S.P. and Cole,M.D. (1982) *Cell*, **31**, 443-452.

Stanton,L.W., Watt,R. and Marcu,K.B. (1983) *Nature*, **303**, 401-406.

Tabin,C.J., Bradley,S.M., Bargmann,C.I., Weinberg,R.A., Papageorge, A.G., Scolnick,E.M., Dhar,R., Lowy,D.R. and Chang,E.H. (1982) *Nature*, **300**, 143-149.

Taparowsky,E., Suard,Y., Fasano,O., Shimizu,K., Goldfarb,M. and Wigler, M. (1982) *Nature*, **300**, 762-765.

Taub,R., Kirsch,I., Morton,C., Lenoir,G., Swan,D., Tronick,S., Aaronson, S. and Leder,P. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 7837-7841.

Watson,D.K., Psallidopoulous,M.C., Samuel,K.P., Dalla-Favera,R. and Papas,T.S. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 3642-3645.

Watt,R., Stanton,L.W., Marcu,K.B., Gallo,R.C., Croce,C.M. and Rovera, G. (1983) *Nature*, **303**, 725-728.

Westin,E.H., Wrong-Staal,F., Gelmann,E.P., Dalla-Favera,R., Papas,T.S., Lautenberger,J.A., Eva,A., Reddy,E.P., Tronick,S.R., Aaronson,S.A. and Gallo,R.C. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 2490-2494.

Received on 13 October 1983