J. Clin. Chem. Clin. Biochem.
Vol. 26, 1988, pp. 195–200


# Evaluation of Random cDNA Clones as Probes for Human Restriction Fragment Length Polymorphisms

By *Cornelia Wappenschmidt, Miyoko Higuchi, C. Meisen*

*Institut für Strahlenbiologie der Universität Bonn, Bonn*

*A. J. Driesel*

*DECHEMA-Institut, Frankfurt/Main*

*K.-H. Grzeschik*

*Institut für Humangenetik, Münster* and

*K. Olek*

*Institut für Strahlenbiologie der Universität Bonn, Bonn*

(Received August 10, 1987/January 22, 1988)

**Summary:** We constructed two human cDNA libraries and selected clones hybridizing with more than five fragments of digested genomic DNA. We assume that these cDNAs detect sequences belonging to gene families. Compared with cDNAs derived from mRNAs of other tissues, the cDNAs of lymphocytes contained a higher proportion of these selected species of cDNA. We assume that these extra cDNAs are tissue-specific. In parallel tests, cDNAs belonging to gene families detected more restriction fragment length polymorphisms than did genomic probes, due to the larger number of restriction sites that can be checked using one probe. However, the chromosomal assignment of these polymorphisms often proved to be very difficult. In addition, we noticed that the mean length of EcoRI fragments hybridizing with our cDNAs is greater than the mean length of fragments hybridizing with randomly chosen genomic probes, possibly due to methylation connected with the inactivation of related active gene sequences.

## Introduction

Restriction fragment length polymorphisms (RFLPs) are potentially useful genetic markers. They can be used to construct a complete genetic linkage map of the human genome, and those markers located close to a mutated locus causing a monogenic disease allow antenatal diagnosis of the hereditary disorder even when the fundamental biochemical defect is not known (1, 2). Thus there is a great need for a large number of DNA probes detecting restriction fragment length polymorphisms.

Putative, suitable probes may be isolated from genomic and cDNA libraries. In comparison with cDNA clones, genomic clones normally contain much longer inserts, suggesting that more recognition sequences may be tested. However, most of the longer inserts are likely to contain repetitive elements, which constitute approximately 30% of the human genome and often alternate with unique sequences. Repetitive sequences of about 300 base pairs in length, interspersed with single copy sequences of approximately 2000 base pairs, are very abundant. According to *Schmid*

& *Deininger* (3) about 52% of the human genome consist of these 'short-period interspersions'. Therefore it is not surprising that up to two-thirds of genomic DNA clones are found to contain repetitive sequences. DNAs of these clones hybridize with numerous fragments from many different regions of the genome and thereby often show hybridization patterns too complex to be interpreted, thus being unusable for restriction fragment length polymorphism studies. Thus, in order to use genomic libraries as a source of probes defining polymorphic loci, it is first necessary to screen out clones containing repetitive elements.

In contrast, cDNA clones mostly represent sequences present in only one or a few copies per haploid genome. *Kurnit* et al. (4) found that approximately 7% of their cDNA clones derived from foetal tissues hybridized strongly to human DNA enriched for repetitive sequences. Thus a screening for single copy probes is generally unnecessary. But since transcribed sequences are more highly conserved than untranscribed DNA regions during evolution, fewer polymorphic sites are to be expected within cDNAs. Another disadvantage is the small insert size, although cDNAs may cover more than one exon thus hybridizing to more than one genomic fragment.

These disadvantages may be compensated using cDNAs that hybridize to several different fragments. In a study of randomly chosen cDNA clones, *Helentjaris & Gesteland* (5) noticed that probes showing numerous hybridization bands revealed several different restriction fragment length polymorphisms.

To check the generality of their findings we constructed two cDNA libraries, and randomly chosen clones were tested with regard to the number of hybridizing fragments. Those probes showing various hybridization bands were investigated for their usefulness in detecting restriction fragment length polymorphisms. Analogous polymorphism studies were carried out using genomic probes.

## Materials and Methods

A human genomic library was constructed using randomly sheared DNA with an average fragment length of 5000 base pairs. Prior to insertion into the PstI site of pBR322, fragments were digested with PstI.

cDNA libraries were prepared using mRNAs from either human lymphocyte cell cultures or sweat gland epithelial cell cultures. The lymphocyte cDNAs were cloned into the PstI site of pUR291 (6) using standard procedures (7), whereas the sweat gland epithelium cDNA library was constructed according to the method described by *Okayama & Berg* (8) using the pcD vector. These libraries were screened for nonrepetitive clones containing inserts of at least 350 base pairs.

Human genomic DNAs were isolated from leukocytes and digested to completion with various restriction endonucleases. The resulting fragments were separated by agarose gel electrophoresis and transferred to nitrocellulose filters using standard procedures (7).

Inserts of suitable clones were isolated using DEAE cellulose and subsequently labeled to a specific activity of approximately $10^8$ counts/min per µg.

Hybridizations were carried out in aqueous formamide, volume fraction 0.5, 5 × SSPE (0.9 mol/l NaCl, 50 mmol/l $NaH_2PO_4$, 5 mmol/l EDTA, pH 7.4), 5 × *Denhardt*'s (1 g/l bovine serum albumin, 1 g/l ficoll, 1 g/l polyvinylpyrrolidone), 50 mg/l carrier DNA at 42 °C.

After hybridization filters were washed in 0.1 × SSC (15 mmol/l NaCl, 1.5 mmol/l sodium citrate, pH 7), 1 g/l SDS at 65 °C, air dried and exposed to X-ray films at −70 °C using two intensifying screens.

Hybridizing restriction fragments were assigned to human chromosomes using a panel of human-rodent somatic cell hybrids.

## Results

Our genomic clones were screened for single-copy sequences, and 50% (25 of 50) were found to be free of repetitive elements. These clones (average insert size 2000 base pairs) were tested for their usefulness as detectors of restriction fragment length polymorphisms by screening DNAs from nine unrelated Caucasians, using at least five restriction endonucleases.

Seven of the probes examined (28%) identified eight restriction fragment length polymorphisms (tab. 1).

This fraction is consistent with previous results of other authors e. g. *Schumm* et al. (9), indicating that about one third of genomic single-copy clones detect restriction fragment length polymorphisms.

*Schumm* et al. (9) also noticed that about 1/2 of their polymorphic probes detected at least three alleles. We could not confirm these results using our genomic probes. None of them detected more than two alleles. This might be explained by the comparatively small number of probes tested.

All clones except one revealed only one or two bands, no matter which restriction enzyme was used to digest the DNA; in the exceptional case mentioned we always detected 4−6 bands. The chromosomal assignment of the identified DNA fragments did not raise difficulties. All hybridization bands detectable with a single clone were localised on the same chromosome.

Screening of our cDNA libraries for the presence of repetitive sequences revealed that at most 7% of the clones contain repetitive regions, thus showing hybridization patterns too complex to be analysed. This value is comparable to that obtainable for cDNAs

Tab. 1. Genomic probes identifying restriction fragment length polymorphisms (RFLPs), chromosomal assignment of polymorphic loci and list of all restriction enzymes tested

| Probe | Polymorphic for | Chromosomal localisation of polymorphic bands | Not polymorphic for |
|---|---|---|---|
| DR6[15] | HindIII | 9 | BamHI, BglII, EcoRI, EcoRV, MspI, TaqI |
| | SstI | 9 | |
| DR10[15] | SstI | 1 | BamHI, BglII, EcoRI, EcoRV, HindIII, MspI, TaqI |
| DR58 | EcoRV | 11 | BamHI, BglII, EcoRI, HindIII, MspI, SstI, TaqI |
| DR78[15] | BglII | 1 | BamHI, EcoRI, EcoRV, HindIII, MspI, SstI, TaqI |
| DR82 | MspI | 3 | BamHI, BglII, EcoRI, EcoRV, HindIII, SstI, TaqI |
| DR110 | MspI | 19 | BglII, EcoRI, PvuII, XbaI |
| DR274 | BamHI | N.D. | BglII, EcoRI, HindIII, MspI, PvuII, TaqI, XbaI |

N.D. — not determined

derived from mRNAs of human foetal tissues. *Kurnit* et al. (4) examined cDNAs of this origin and also found that approximately 7% hybridized strongly to human repetitive sequences.

Subsequently, 43 nonrepetitive cDNAs with an average insert length of 450 base pairs were hybridized to EcoRI-restricted human DNA. Based on the number of detectable fragments, three groups of probes were separated:
firstly probes detecting at most two fragments (I),
secondly those hybridizing with up to five fragments (II) and
thirdly cDNAs showing considerably more than five clearly separable bands (III) (fig. 1).

The distributions of these three classes among clones of the two different cDNA libraries are shown in table 2.

There were no great differences in the number of detectable fragments when other restriction endonucleases were used, i.e. membership of the mentioned groups was not influenced by the choice of the restriction enzyme.

cDNAs of group III should be most useful for the detection of restriction fragment length polymorphisms, since the large number of clearly separable fragments means that more potentially polymorphic restriction sites can be examined.

Consequently we tested 13 cDNAs belonging to this group, using at least six restriction enzymes. In fact eight of these probes (62%) detected at least one polymorphism (tab. 4), while an additional cDNA revealed a variant.

In most instances there are two alleles at the polymorphic locus. Only two probes showed more variability detecting three polymorphic alleles.

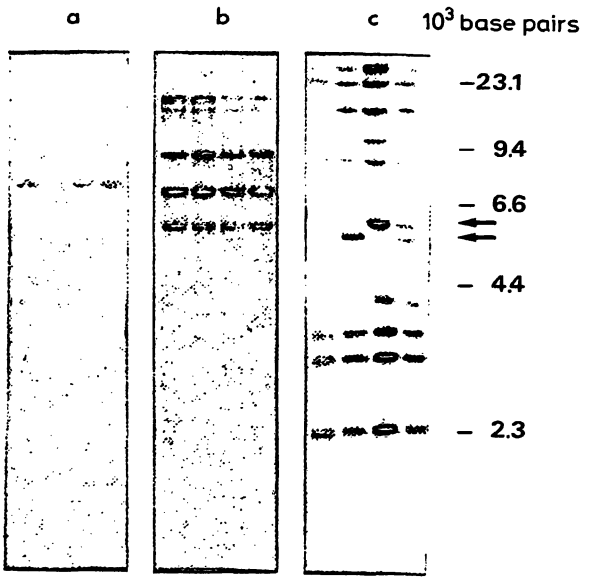


Fig. 1. Representatives of the cDNA categories separable with regard to the number of detectable fragments
a — EcoRI hybridization pattern of probe cCW159
b — EcoRI hybridization pattern of probe cCW26
c — BglII hybridization pattern of probe OL202 with polymorphic bands at 5200 base pairs and 5000 base pairs (indicated by arrows)
The probes cCW159 and cCW26 are derived from sweat gland epithelial cells; the clone OL202 contains a lymphocyte cDNA.

Tab. 2. Classification of cDNA probes according to the number of detectable fragments
I — maximum of two hybridizing fragments
II — from three to five hybridizing fragments
III — more than five hybridizing fragments

| cDNAs derived from | I | II | III |
|---|---|---|---|
| Sweat gland epithelium | 9 (69%) | 2 (15%) | 2 (15%) |
| Lymphocytes | 10 (33%) | 8 (27%) | 12 (40%) |

Using somatic cell hybrids cross-hybridizations of DNA from the rodent parental cell lines with the human probes obstructed the chromosomal assignment. Another difficulty resulted from the finding that not all fragments which were identified by one and the same cDNA clone originated from the same chromosome, i. e. hybridization signals corresponding to the individual fragments were obtained using DNAs of hybrid cell lines containing diverging portions of the human genome. Some examples of the chromosomal assignment of the various fragments detected by individual cDNA probes are given in table 3. It was ascertained that none of the cDNAs detected fragments derived from one chromosome exclusively. Thus in situ hybridizations or localisation studies using other restriction enzymes could only be of limited use.

Owing to these problems, only six of the detectable restriction fragment length polymorphisms could be assigned to specific chromosomes (tab. 4).

Tab. 3. Examples of cDNA probes each hybridizing to fragments of various chromosomal origin

| Probe | Chromosomal assignment of various hybridizing fragments |
|---|---|
| OL11 | 3, 6, 12, 18 |
| OL37 | 1, 5, 7, 10, 12, 14, 18, X |
| OL38 | 6, 9, 13 |
| OL43 | 5, 6, 8, 14 |
| OL202 | 5, 12, 13, 17 |

In addition to the large number of detectable fragments, we noticed that most of the fragments hybridizing with our cDNAs show considerable lengths, the mean value amounting to 9380 base pairs using the enzyme EcoRI. In comparison, EcoRI fragments detected with randomly chosen genomic probes show a mean value of only 5360 base pairs.

The finding that the mean value for genomic probes also exceeds the calculated average fragment length obtainable with EcoRI (3200–3500 base pairs) (10, 11) is probably explained by the fact that the usual test conditions do not allow the detection of very short fragments ($\ll$ 500 base pairs).

Normal distributions were obtained using the logarithms of the fragment lengths observable with cDNAs or genomic clones respectively (fig. 2). The difference of the mean values proved to be statistically significant ($P \leqslant 0.0001$) by *Student*'s t-test.

## Discussion

The multiplicity of fragments hybridizing with each of the tested cDNA probes indicated the presence of several genomic regions showing strong homologies with the cloned sequences. With regard to the small average length of our probes (450 base pairs) this might be explained by the assumption that these cDNAs were derived from mRNAs representing several exons. Thus a cDNA probe would cover some exons and since intervening sequences are present, a large number of fragments would be identified.

Tab. 4. cDNA probes detecting restriction fragment length polymorphisms (RFLPs), chromosomal assignment of polymorphic loci and list of all restriction enzymes tested

| Probe | Polymorphic for | Chromosomal localisation of polymorphic bands | Not polymorphic for |
|---|---|---|---|
| OL11[16] | EcoRI | 12 | BamHI, BglII, HindIII, PvuII, XbaI |
| | MspI | N. L. | |
| OL37[17] | EcoRI | N. L. | BamHI, BglII, EcoRV, HindIII, MspI, SstI, TaqI |
| OL38[17] | EcoRI-a | 13 | BamHI, BglII, EcoRV, HindIII, MspI, SstI, TaqI |
| | EcoRI-b | N. L. | |
| | EcoRI-c | N. L. | |
| OL43[17] | EcoRI | 6 | BamHI, BglII, EcoRV, MspI, SstI, TaqI |
| | HindIII | N. L. | |
| OL167[18] | EcoRI | 21 | BamHI, BglII, HindIII, MspI, PvuII, XbaI |
| OL181 | HindIII | N. L. | BamHI, BglII, EcoRI, EcoRV, MspI, SstI, TaqI |
| OL202[19] | BglII | 12 | BamHI, HindIII, MspI, PvuII, TaqI, XbaI |
| cCW147[20] | EcoRI | N. L. | BamHI, BglII, MspI, XbaI |
| | HindIII | 5 | |

N. L. – not localisable using human-rodent somatic cell hybrids

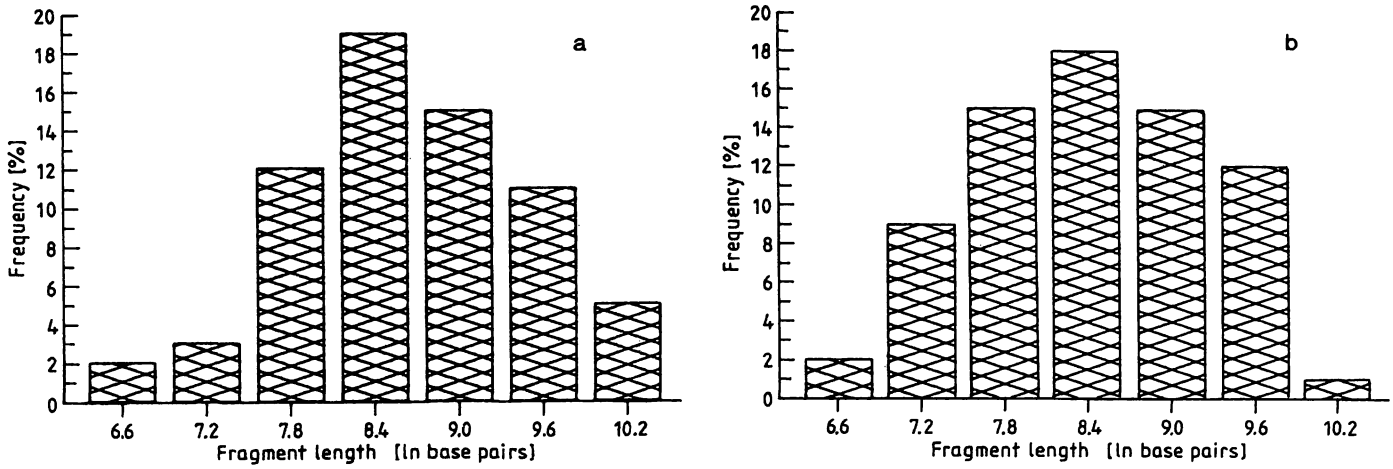


Fig. 2. Normal distributions obtainable using logarithms of the length of EcoRI fragments detectable with cDNA probes (a) or genomic probes (b).

In this situation, all hybridizing sequences must be located on the same chromosome. As already mentioned, however, the various fragments detected by one cDNA could not be attributed to a single chromosome. The different fragments containing sequences homologous to one cDNA probe could rather be assigned to several chromosomes; at the most one probe detected sequences belonging to eight distinct human chromosomes (tab. 3).

Thus it has to be assumed that regions homologous to the cDNAs tested belong to families of related sequences distributed throughout the genome. Such gene families have already been described for several eukaryotic genes (globin genes, immunoglobulin genes, genes for rRNA, actin genes etc.). All members of a gene family are derived from a single ancestral gene. Related genes could have arisen by gene duplication followed by mutations causing the sequence variations. Some of these mutations (e. g. frameshift) may have caused an inactivation and led to the formation of nonfunctional pseudogenes that cannot be translated into functional proteins. Diverging chromosomal locations of related sequences may be due to translocation events during evolution.

Besides those pseudogenes that show the exon-intron arrangement of their productive counterparts, another category of pseudogenes lacking the intervening sequences of their functional relatives has been discovered. Such 'processed' pseudogenes have been found in a number of different gene families and seem to be very common. One of the characteristics of many of these pseudogenes is an oligo(A)tract immediately 3′ to the point at which their homology with the functional gene ceases. It is therefore supposed that they are derived in some way from mRNA and that they are the result of the integration of cDNA copies into the genome.

It has been found that processed pseudogenes and their functional relatives are not syntenic, i. e. they are not located on the same chromosome, whereas the majority of pseudogenes showing an exon-intron arrangement are linked to their productive counterparts (12). In addition it has been found that functional genes belonging to one gene family are located on at most two or three chromosomes, often within clusters. Therefore it has to be assumed that translocation events have been of minor importance to the distribution of related sequences throughout the genome.

It is therefore obvious that at least some of the members of the gene families detected with our cDNA probes represent processed pseudogenes, which have strong homology to the cDNA but are not located on the same chromosome(s) as the functional gene(s). This explanation for the large number of identified chromosomes seems to be more probable than an accumulation of translocation events following gene duplications.

A remarkable example of a gene family with several processed pseudogenes is the human argininosuccinate synthetase gene family, which is composed of one functional gene and 14 processed pseudogenes assigned to 11 different chromosomes (13).

The proportion of cDNA probes that identify gene families (tab. 2, category III) was higher in lymphocyte cDNAs than in cDNAs derived from sweat gland epithelial cells. As the value found for the latter library is consistent with the results obtainable for human brain or placental cDNAs (5) a proportion of about 15% is assumed to be generally accepted for most cDNA libraries. We assume that the considerably higher percentage found for lymphocyte cDNAs is tissue specific. Possibly the immunoglobulin supergene family (immunoglobulins, T-cell receptors, major

histocompatibility complex) is of importance here. Extensive localisation studies should prove whether accumulations of corresponding chromosomal regions (e. g. chromosomes 2, 14 and 22 for functional immunoglobulin genes) are detectable.

The finding that cDNAs identifying gene families hybridize with DNA fragments of longer average size than randomly chosen genomic probes may be interpreted as the result of a nonrandom distribution of EcoRI sites, but it is also conceivable that some recognition sequences are concealed by methylation of certain bases. It is known that $N^6$-methyladenine or 5-methylcytosine residues at the indicated positions within the recognition sequence GAA*TTC* inhibit the enzymatic activity of EcoRI (14). This loss of cleavage sites caused by methylation leads to the occurrence of longer DNA fragments. As gene families often represent sequences that accomplish the same task in different tissues or during different developmental periods it is conceivable that methylation is used to inhibit transcription of other active genes.

However, regarding our previous comments, only a few fragments should be derived from active genes, whereas several other fragments should represent nonfunctional relatives whose methylated conditions are still unknown.

Methylation may also be responsible for the occurrence of some restriction fragment length polymorphisms, at least indirectly, due to relatively frequent transition of methylated cytosines in CpG dimers.

Comparing the restriction fragment length polymorphism output of cDNAs and genomic probes, it is remarkable that the percentage of cDNAs of category III is considerably higher (62%) than that of single copy genomic probes (28%). However, taking into account the number of restriction sites tested, more restriction sites have to be screened to detect one polymorphic site when cDNAs were used. This might be due to the fact that cDNAs represent sequences which are more highly conserved during evolution, thus revealing fewer polymorphisms. The higher portion of cDNA probes detecting restriction fragment length polymorphisms is the result of the large number of restriction sites that can be checked simultaneously using one probe.

Although we found several restriction fragment length polymorphisms using our cDNA probes, only a few could be assigned to specific human chromosomes. This constitutes a major problem in the application of these cDNAs for the detection of restriction fragment length polymorphisms, because the chromosomal assignment of restriction fragment length polymorphisms is fundamental to their usefulness as genetic markers.

**Acknowledgement**

This study was supported by a grant from "Bundesminister für Forschung und Technologie" and Boehringer Mannheim to A. J. D., K.-H. G. and K. O. for the joint project PTB 03-8535-8.

**References**

1. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. (1980) Am. J. Hum. Genet. *32*, 314–331.
2. Davies, K. E. (1981) Hum. Genet. *58*, 351–357.
3. Schmid, C. W. & Deininger, P. L. (1975) Cell *6*, 345–358.
4. Kurnit, D. M., Wentworth, B. M., de Long, L. & Villa-Komaroff, L. (1982) Cytogenet. Cell Genet. *34*, 193–203.
5. Helentjaris, T. & Gesteland, R. (1983) J. Mol. Appl. Genet. *2*, 237–247.
6. Rüther, U. & Müller-Hill, B. (1983) Embo J. *2*, 1791–1794.
7. Maniatis, T., Fritsch, E. F. & Sambrook, J. (eds) (1982) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
8. Okayama, H. & Berg, P. (1983) Mol. Cell. Biol. *3*, 280–289.
9. Schumm, J., Knowlton, R., Braman, J., Barker, D., Vovis, G., Akots, G., Brown, V., Gravius, T., Helms, C., Hsiao, K., Rediker, K., Thurston, J., Botstein, D. & Donis-Keller, H. (1985) HGM8, Cytogenet. Cell Genet. *40*, 739.
10. Bastié-Sigeac, F. & Lucotte, G. (1983) Hum. Genet. *63*, 162–165.
11. Drmanac, R., Petrovic, N., Glisin, V. & Crkvenjakov, R. (1986) Nucl. Acids Res. *14*, 4691–4692.
12. Vanin, E. F. (1984) Biochim. Biophys. Acta *782*, 231–241.
13. Su, T. S., Nussbaum, R. L., Airhart, S., Ledbetter, D. H., Mohandas, T., O'Brien, W. E. & Beaudet, A. L. (1984) Am. J. Hum. Genet. *36*, 954–964.
14. McClelland, M. & Nelson, M. (1985) Nucl. Acids Res. *13* (Suppl.), r201–r207.
15. Icking, D., Jandel, A.-S., Kabisch, R., Driesel, A. J., Olek, K. & Grzeschik, K.-H. (1985) HGM8, Cytogenet. Cell Genet. *40*, 659.
16. Wappenschmidt, C., Higuchi, M., Driesel, A. J., Grzeschik, K.-H. & Olek, K. (1987) Nucl. Acids Res. *15*, 2398.
17. Jandel, A.-S., Icking, D., Kabisch, R., Olek, K., Driesel, A. & Grzeschik, K.-H. (1985) HGM 8, Cytogenet. Cell Genet. *40*, 660–661.
18. Wappenschmidt, C., Higuchi, M., Driesel, A. J., Grzeschik, K.-H. & Olek, K. (1987), Nucl. Acids Res. *15*, 7658.
19. Wappenschmidt, C., Higuchi, M., Driesel, A. J., Grzeschik, K.-H. & Olek, K. (1986) Nucl. Acids Res. *14*, 9221.
20. Wappenschmidt, C., Meisen, C., Driesel, A. J., Grzeschik, K.-H. & Olek, K. (1987) Nucl. Acids Res. *15*, 2399.

Dr. Cornelia Wappenschmidt
Institut für Strahlenbiologie
Sigmund-Freud-Str. 15
D-5300 Bonn