Analysis of the regions flanking the human insulin gene and sequence of an Alu family member

Graeme I.Bell, Raymond Pictet and William J.Rutter

Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, CA 94143, USA

Received 21 July 1980

ABSTRACT

The regions around the human insulin gene have been studied by heteroduplex, hybridization and sequence analysis. These studies indicated that there is a region of heterogeneous length located approximately 700 bp before the 5' end of the gene; and that the 19 kb of cloned DNA which includes the 1430 bp insulin gene as well as 5650 bp before and 11,500 bp after the gene is single copy sequence except for 500 bp located 6000 bp from the 3' end of the gene. This 500 bp segment contains a member of the Alu family of dispersed middle repetitive sequences as well as another less highly repeated homopolymeric segment. The sequence of this region was determined. This Alu repeat is bordered by 19 bp direct repeats and also contains an 83 bp sequence which is present twice. The regions flanking the human and rat I insulin genes were compared by heteroduplex analysis to localize homologous sequences in the flanking regions which could be involved in the regulation of insulin biosynthesis. The homology between the two genes is restricted to the region encoding preproinsulin and a short region of approximately 60 bp flanking the 5' side of the genes.

INTRODUCTION

The factors which govern the cell specific and temporal appearance of insulin during development as well as those which modulate its biosynthesis at the genetic level are unknown, but presumably they operate on the DNA near the gene. We have isolated a 19 kb segment of human chromosome 11 which includes the 1430 bp insulin gene, and have determined the sequence of the gene and adjacent regions (1,2). By comparing the sequence of the regions flanking the human insulin gene with those for the corresponding regions of the rat insulin I and II genes (3,4), several conserved and potential regulatory sequences were identified. However the function(s) of the remainder of the sequence around the human insulin gene and their role, if any, in the regulation of insulin biosynthesis is undetermined. This approximately 17.5 kb of DNA could include other insulin specific regulatory sequences, non-expressed spacer sequences, as well as other genes.

In this paper we extend the analysis of the regions flanking the human insulin gene. We have determined that there is a region approximately 700 bp from the 5' end of the gene which is polymorphic in length. We have compared human and rat genomic DNA fragments containing the insulin gene and determined that the homology in the 5' flanking region does not extend beyond 60 bp from the gene. In addition we have searched for repetitive sequences in the vicinity of the insulin gene. Repeated sequences account for approximately 20% of the human genome (5) and at least five families have been identified (6-10). We have detected a single repetitive sequence of 313 bp within the 19 kb DNA segment containing the insulin gene located approximately 6 kb from the 3' end of the gene. The sequence indicates that it is a member of the Alu family of dispersed middle repetitive sequences (11) and is flanked by 19 bp direct repeats. The partial sequences of two other members of this family obtained by cloning middle repetitive sequences and a consensus sequence for this family have been reported (11), however the sequence presented here is the first to describe the unique flanking sequences as well.

MATERIALS AND METHODS

DNA Samples

The isolation and analysis of human insulin genomic DNA fragments from a fetal liver DNA library (library 1 of Fritsch et al. (12) obtained from Dr. T. Maniatis) has been described previously. The isolation and analysis of the rat genomic DNA fragment containing the insulin I gene is described in Cordell et al. (3).

Preparation of ^{32}P-Labeled DNA

DNA was labeled by nick translation essentially as described by Rigby et al. (13). The specific activities obtained were 1-4x10^{8} cpm/μg.

Electrophoresis, Blotting and Hybridization

After digestion, DNA fragments were separated by electrophoresis in a vertical agarose slab gel (10x14x0.3 cm) in 40mM Tris/acetate pH 8.1, 20mM sodium acetate and 2mM disodium ethylenediaminetetraacetic acid. Fragments of HindIII-digested lambda DNA and HaeIII-digested ØX174 were used as molecular weight markers. Composite acrylamide:agarose gels with the reversible cross-linking reagent, N,N'-diallyltartardiamide, (14), were prepared in the same buffer as for agarose gels. DNA was transferred from gels to Schleicher and Schuell nitrocellulose filters (BA85) as described by Southern (15). Composite gels were soaked in 2% periodic acid for 30 min at 37°C before transfer. After transfer, filters were baked in vacuo for 2 hr at 80°C. Hybridization was for 24 hr at 42°C in the buffer described by Wahl et al. (16) which includes 10% dextran sulfate. ^{32}P-labeled DNA was present at 2.5×10^5 cpm/ml. Filters were washed after hybridization as described by Wahl et al. (16) and then exposed to Kodak XR-2 X-ray film with Dupont Lightning Plus Intensifying Screen at -76°C.

DNA Sequence Analysis

DNA sequencing was with the procedure of Maxam and Gilbert (17) as described previously (1).

RESULTS

Length Heterogeneity in Region Flanking 5' end of Human Insulin Gene

Figure 1 shows the organization of the 19 kb DNA segment containing the insulin gene and flanking sequences. As discussed previously (1) the insulin genes in λH-1 and λH-2 are alleles because of sequence differences in the gene region encoding the 3' untranslated portion of the

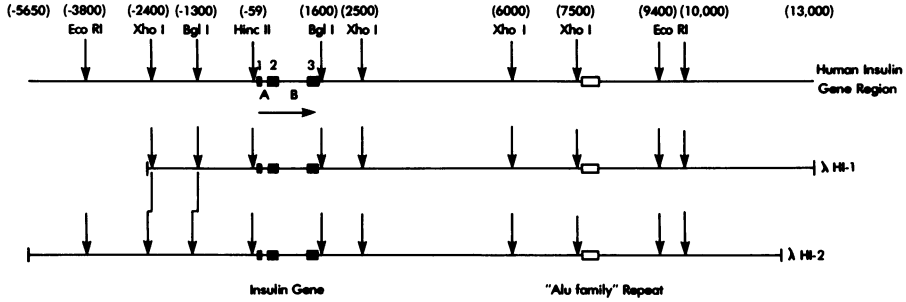

FIGURE 1. A map of the human insulin gene. The map of relevant restriction enzyme sites is presented in the top line. The coordinates are in bp relative to the 5' end of preproinsulin mRNA (position 1). The dark boxes (1-3) are the mRNA coding regions. A and B are intervening sequences. The arrow indicates the direction of transcription. The position of the Alu family repeated sequence is indicated by the open box. The λCh4A-cloned human DNA fragments are indicated below the map. As described in the text λHI-2 contains a 100 bp insertion located approximately at position -700. We have adjusted the sizes of some of the fragments described previously (1) based upon more extensive determinations.

mRNA. The analysis of these two cloned fragments also showed that they had very similar restriction maps. However subsequent experiments revealed that the XhoI subfragment which contained the insulin gene was 100 bp larger in λHI-2. A heteroduplex formed between the corresponding XhoI subfragments (coordinates approximately -2400 and 2500) of the two cloned DNA segments indicated an insertion-deletion loop (Fig. 2). Further restriction mapping and hybridization analyses established that the insertion was centered approximately 700 bp before the 5' end of the insulin gene of λHI-1 (data not shown). This does not appear to be a cloning artifact (12,18) since we have observed that EcoRI digestion of some individuals (8/15) produces two fragments of different sizes which contain insulin gene sequences (for example, see Fig. 4A, lane 2). Each fragment corresponds to one of the parental insulin genes and adjacent regions, since there are no EcoRI sites within the cloned insulin genes. If there were an EcoRI site within one of the insulin genes, EcoRI

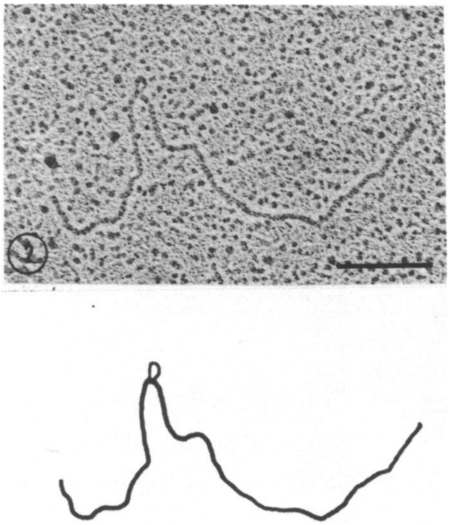

FIGURE 2. Heteroduplex of the insulin gene containing XhoI fragment from λHI-1 and λHI-2. The XhoI fragments (coordinates -2400 to 2500 (Fig. 1)) were prepared by electrophoresis. 0.1 μg of each fragment was mixed and denatured in 100 μl of 0.1M NaOH, 20mM EDTA for 10 min at room temperature. The solution was neutralized by addition of 20 μl of 1M Tris HCl pH 7.5 and formamide was added to 50%. Hybridization was for 4 hr at 25°C. 50 μl aliquots were diluted four fold into the spreading solution (50% formamide, 10mM Tris, 1mM EDTA, 50 μg/ml cytochrome), and were spread on the hypophase (15% formamide, 10mM Tris HCl, pH 8.0, 1mM EDTA). The duplex regions are 3360 bp and 1540 bp. The bar represents 0.2 μ.

digestion would generate three insulin gene containing fragments and the sum of the sizes of the two smaller fragments should equal the larger, approximately 13 kb. This has never been observed. This difference is also not due to a restriction site polymorphism in a region outside the gene at least in three of these individuals. Digestion of their DNA with BglII (coordinates -168 and 9340, Fig. 1), generates a single insulin gene containing fragment whereas BglI and XhoI (Fig. 1) generate two fragments as expected since they span the insertion region. In these three individuals the insertion is between the BglI site at -1300 and a PvuII site at -259. The sizes of the insertions are of the order of 1000-2000 bp (insertions of less than 250 bp probably could not be detected) which are much larger than observed in λHI-2 and it remains to be determined if much smaller insertions also occur. We have not detected any other size differences of corresponding subfragments of λHI-1 and λHI-2 between coordinates -2400 and 10,000.

Comparison of Sequences Flanking Human and Rat I Insulin Genes

A comparison of the sequences of human and rat insulin genes demonstrates that the genes were homologous within the insulin coding regions but that the intervening sequences and those following the codon for translation termination were divergent (1). Furthermore an examination of 106 bp before the genes indicated islands of homology interspersed within nonhomologous regions. These included the AT-rich region, the Hogness box, centered 27 bp before the gene as well as another 80 bp before the gene. Since the sequence was not extensive enough to allow a more complete analysis, we compared the sequences flanking the human and rat I insulin genes indirectly by analyzing heteroduplexes formed between the cloned genes (Fig. 3). There are five regions of homology within the heteroduplexes: at each end, corresponding to the pBR322 sequences, and three (A,B, and C) near the middle of the DNA molecules, the region encoding insulin. The arrangement of the duplex regions indicated that the polarity of the genes is 5'-A,B,C-3'. The region of homology, A (Fig. 3), corresponds to the region from the 5' boundary of intervening sequence 1 and extending into the 5' flanking region of the gene, B (Fig. 3), from near the 3' boundary of intervening sequence 1 to the 5' boundary of intervening sequence 2, and finally C (Fig. 3), from the 3' boundary of intervening sequence 2 to the translation termination codon. A and B (Fig. 3) are separated by nonhomologous intervening sequence 1 and B and C by intervening sequence 2 (rat I insulin gene lacks this

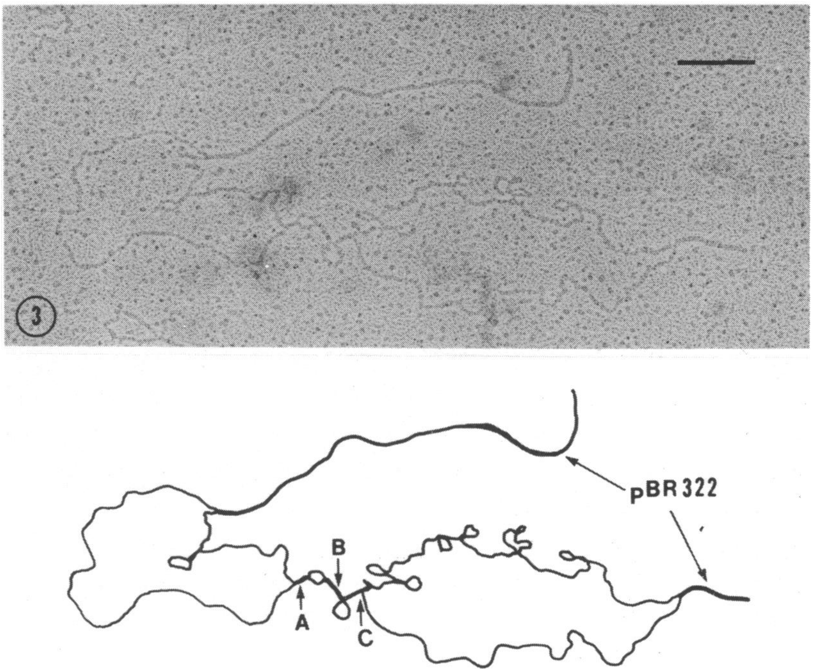

FIGURE 3. Heteroduplex of the human and rat I insulin genes and flanking regions. The 9.4 kb EcoRI fragment containing the rat I insulin gene (3) and the 13.2 kb EcoRI fragment containing the human insulin gene of λHI-2 were cloned in EcoRI site of pBR322. The recombinant plasmids were linearized by cleavage of the SalI site within pBR322. Heteroduplexes were formed as described in Fig. 2. The double stranded DNA region on each end of the molecule represents the pBR322 sequence on each side of the insert. A,B and C designate the regions of homology between the human and rat I insulin genes. The single stranded region between A and B represents intervening sequence 1. The loop between B and C represents intervening sequence 2 which is only present in the human insulin gene. The bar is 0.2 μ.

intervening sequence and therefore there is an insertion loop from the human DNA in this region of the heteroduplex). The size of region A, relative to B and C, 234 bp and 154 bp respectively, whose sizes are precisely known, from sequence analysis (1) is 100-120 bp. This is the same as expected from the sequence comparison; therefore the homology at the 5' ends of the genes does not extend outside the region sequenced. The absence of other hybrid regions indicates that there are no regions of extensive homology between the human and rat I insulin genes besides those previously determined. We verified the heteroduplex analysis by hybridizing the labeled rat insulin gene containing EcoRI fragment to subfragments of the 13.2 kb EcoRI fragment of λHI-2. There was no detectable homology outside the region of the insulin gene (data not shown).

Location of Repetitive Sequences Flanking Human Insulin Gene

When we initially used the EcoRI fragment (coordinates -3800 to 9400 (Fig. 1)) containing the insulin gene from λHI-2 as a probe in hybridization to EcoRI digested DNA, the EcoRI fragments containing insulin sequences were evident (Fig. 4A, lane 1) but the background hybridization was high and with longer times the lane was completely exposed. This suggested that this 13 kb EcoRI fragment also contains sequences which are represented many times elsewhere in the genome and are present on EcoRI fragments of all sizes. XhoI digestion of the 13 kb EcoRI fragment generates five fragments of 5.0, 3.5, 1.9, 1.5 and 1.4 kb (Fig. 1). Each of these fragments was isolated, labeled by nick translation and hybridized to EcoRI digested DNA. The 5.0 kb XhoI subfragment (coordinates -2400 to 2500 (Fig. 1)) hybridized only to the two "allelic" EcoRI fragments which contain insulin gene sequences (Fig. 4A, lane 2) and of which it is derived as did the 3.5, 1.5 and 1.4 kb subfragments (coordinates 2500 to 6000, 6000 to 7500 and -3800 to -2400, respectively (Fig. 1)) (data not shown). The cloned insulin cDNA also hybridizes to the same size EcoRI fragments of the DNA preparation as the subfragments of the cloned genes indicating that there is a single insulin gene in the haploid human genome and that the difference in EcoRI fragment length is indicative of a length polymorphism in a region flanking the gene. Moreover the hybridization of the 5.0 kb XhoI subfragment which includes the insulin gene with flanking and intervening sequences to a unique EcoRI fragments indicates that these sequences are single copy. The 1.9 kb XhoI subfragment (coordinates 7500 to 9400 (Fig. 1)) did not hybridize specifically to any EcoRI fragment but did to many different EcoRI fragments (Fig. 4A, lane 3) and thus contains repetitive sequences. This 1.9 kb XhoI subfragment was further digested with SacI which generates three fragments of 935, 542 and 375 bp (coordinates 8420 to 9400, 7500 to 8042, and 8042 to 8420, respectively) and the hybridization analysis was repeated (Fig. 4A, lanes 4-6). The 542 and 375 bp SacI fragments contain repeated sequences.

It is also possible to locate repetitive sequences in isolated

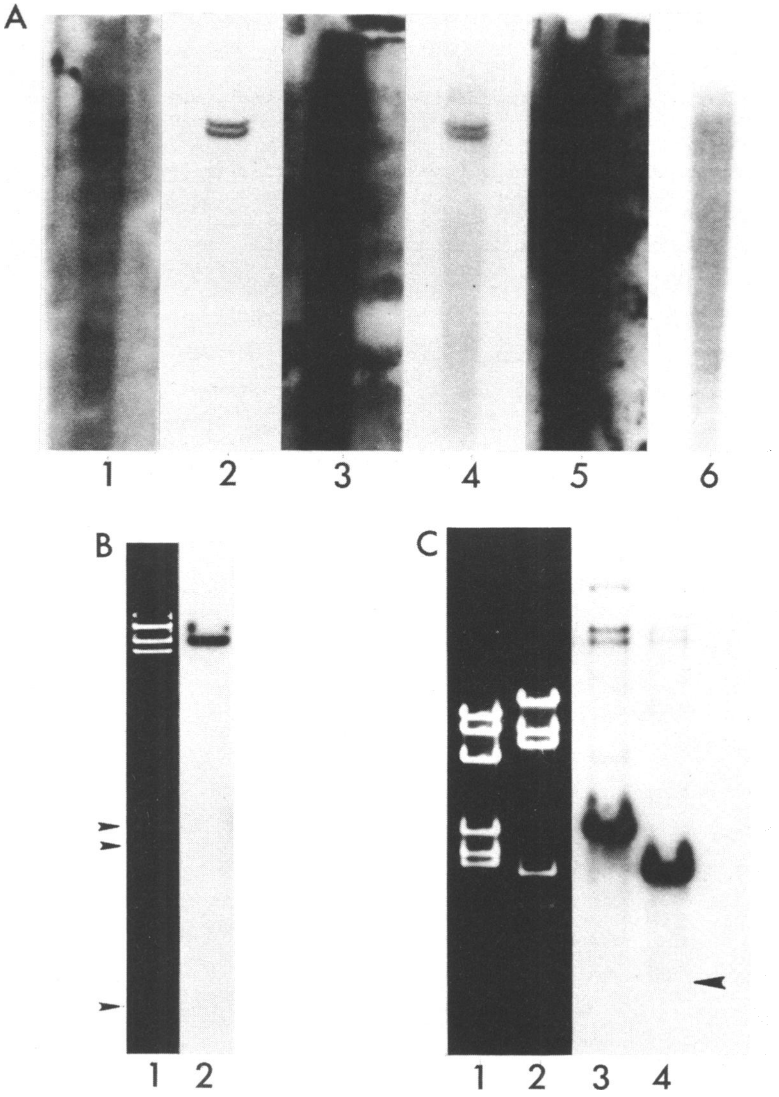

FIGURE 4. Identification of repeated sequences around the human insulin gene. A. Hybridization of human DNA with ^{32}P-labeled restriction fragments. DNA prepared from a male placenta was digested with EcoRI, electrophoresed in a 0.85% agarose gel and transferred to nitrocellulose filters. Fragments of λHI-2, subcloned in pBR322 and then purified by electrophoresis were ^{32}P-labeled and hybridized to the filters. As described in the text, there is a length polymorphism in this DNA preparation in the region between the BglI site at -1300 and a PvuII site at -269 and therefore hybridization is seen to two EcoRI fragments of

approximately 15 and 13 kb. The 15 kb fragment contains a 2 kb insertion not present in the other fragment. The ^{32}P-labeled restriction fragments identified by their coordinates in Fig. 1 are: 1. EcoRI fragment -3800 to 9400. 2. XhoI fragment -2400 to 2500 (this fragment contains a 100 bp sequence in λHI-2 not present in λHI-1 (see text) but identical results were obtained when this fragment was prepared from λHI-1. 3. XhoI-EcoRI fragment 7500 to 9400. 4. SacI-EcoRI fragment 8400 to 9400. 5. XhoI-SacI fragment 7500 to 8040. 6. SacI fragment 8040 to 8420. The extensive hybridization observed in lanes 1,3,5 and 6 is of sequences in the ^{32}P-labeled restriction fragment which are repeated in the human genome and present on EcoRI fragments of all sizes. B. DNA prepared from λHI-2 was digested with EcoRI, electrophoresed in a 0.85% agarose gel, stained with ethidium bromide (left lane) and hybridized with ^{32}P-labeled human DNA (right lane). The fragments sizes (in kb) are 19.8 (lambda arm), 13.2, 10.9 (lambda arm), 2.2, 1.85 and 0.6 (the hybridizing fragment is underlined). Arrowheads indicate the positions of these last three fragments. These were evident in the original photograph. C. The 13.2 kb EcoRI subfragment of λHI-2, subcloned in pBR322, was digested with EcoRI and then XhoI (lanes 1 and 3) or SacI (lanes 2 and 4), electrophoresed in a 0.85% agarose gel, stained with ethidium bromide (left side) and hybridized with ^{32}P-labeled human DNA (right side). The fragment sizes (in kb) are: 1. EcoRI-XhoI digest, 5.0, 4.4 (pBR322), 3.5, 1.9, 1.5 and 1.4 (the hybridizing fragments are underlined). 2. EcoRI-SacI digest, 6.0, 4.4 (pBR322), 4.0, 1.3, 0.9, 0.4 (this fragment was not visible in this stained gel and the region where hybridization to fragments of this size would be expected is indicated by the arrowhead).

genomic DNA segments by hybridizing labeled genomic DNA to immobilized cloned restriction fragments since with standard hybridization conditions only repeated sequences will hybridize (19). This is a rapid and facile procedure since it does not require the isolation of large numbers of DNA fragments. We used this procedure to confirm and extend our initial studies (Fig. 4 and 5). As expected the 13 kb EcoRI fragment of λHI-2 contained repetitive sequences (Fig. 4B). Its 1.9 kb XhoI subfragment hybridized with the genomic DNA probe (Fig. 4C, lane 3) as well as a 1.25 kb SacI subfragment (coordinates 6790 to 8042) (Fig. 4C, lane 4). In contrast, to the results described above, there was no hybridization to a 375 bp SacI fragment. This hybridization analysis of the 1.9 kb XhoI subfragment indicated that the smallest fragment which contains repetitive sequences is the 300 bp XhoI-BglI fragment (coordinates 7500 to 7800) (Fig. 5, lane 2). The two procedures furnished complementary results except in one case. Using isolated restriction fragments as probes, the 375 bp SacI fragment also appeared to contain repetitive sequences (Fig. 4A, lane 6), however this fragment did not hybridize with the nick translated genomic DNA. The reason for this discrepancy is unknown, but as discussed below this fragment contains a 40 bp homo-

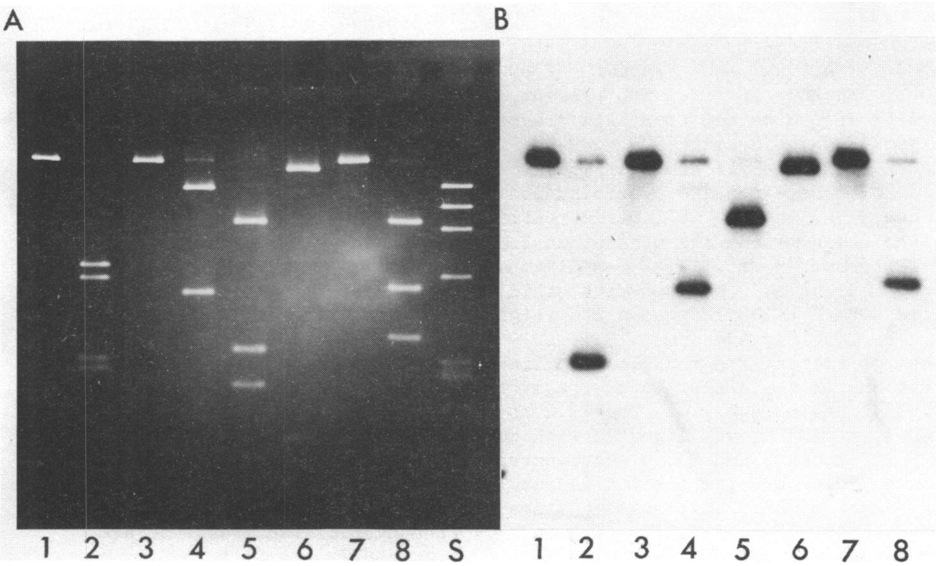

FIGURE 5. Identification of repeated sequences in the 1.9 kb XhoI-EcoRI fragment 7500 to 9400. DNA was digested, electrophoresed in a 4% acrylamide, 0.6% agarose composite gel cross-linked with N,N'diallyltartardiamide, stained with ethidium bromide (A), transferred to nitrocellulose and hybridized with ^{32}P-labeled human DNA (B). The restriction enzymes and fragment sizes are: 1. Undigested DNA, 1900 bp (the hybridizing fragment is underlined). The size of the undigested fragment is 1850-1900 bp. 2. BglI-digested DNA, 650 bp, 600 bp, 318 bp, 300 bp. 3. PstI-digested DNA, 1900 bp (no cleavage sites). 4. PvuII-digested DNA, 1350 bp and 528 bp. 5. AvaI-digested DNA, 900 bp, 355 bp, 270 bp, 160 bp, 150 bp, 29 bp and 15 bp. The latter two fragments determined by sequence. 6. XmaI-digested DNA, 1650 bp, 180 bp and 15 bp (determined by sequence). 7. HaeII-digested DNA, 1900 bp (no sites). 8. SacI-digested DNA, 935 bp, 542 bp and 375 bp. S. HaeIII-digested ØX174 DNA standards.

polymeric sequence which is present at least twice in the 1.9 kb XhoI subfragment. It could also be present elsewhere in the genome but at a lower frequency than the other repetitive sequence and therefore does not hybridize with the labeled genomic DNA under the conditions used.

We also used 2D-Southern blotting (19) to determine if any of the sequences in each of the XhoI subfragments of the 13 kb EcoRI fragment of λHI-2 (see Fig. 1) was present elsewhere within this EcoRI fragment. There was no cross hybridization between any of the XhoI subfragments. Other 2D-Southern blots also indicated no homology between the four EcoRI fragments of λHI-2 as expected from experiments described above (data not shown).

These analyses establish that this 19 kb segment of human DNA contains a repeated sequence of approximately 300 bp and possibly a shorter sequence of approximately 40 bp which is also present elsewhere in the genome but at a lower frequency.

Sequence of Repetitive DNA Element Flanking the Human Insulin Gene

The hybridization analyses described above established the boundaries of a repetitive sequence element centered approximately 6 kb from the 3' end of the human insulin gene. In order to identify the repeated sequence and to determine the nature of the non-repetitive flanking region we sequenced λHI-1 in the vicinity of the repeat. A restriction map for the 1.9 kb XhoI subfragment and the strategy for sequencing this region are presented in Fig. 6. The nucleotide sequence determined for this segment (Fig. 7) indicates that the repeat is a member of the Alu family of dispersed middle repetitive sequences (11). The sequence is limited by almost perfect 19 bp direct repeats (Fig. 7, positions 54-72 and 386-404, Fig. 8). There is only one mismatch in the sequence AAAACAAGCAGGAGAGGCT. If these direct repeats mark the boundaries of the
G
Alu family repetitive sequence, then this member (Fig. 7, positions 73-385) is 313 bp exclusive of terminal repeats which is similar to a size of approximately 300 bp determined for members of this family by reassociation kinetics (9). The hybridization experiments described above placed the boundaries of the repeated sequence near the XhoI site and the BglI site centered around nucleotide 300. There was no detectable

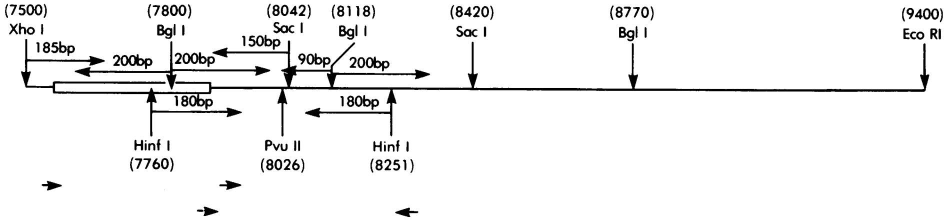

FIGURE 6. Restriction map of the 1.9 kb XhoI-EcoRI fragment containing the repeated sequences and the strategy for determining their primary structure. The coordinates of relevant restriction sites are indicated. The direction and number of bases determined from each labeled restriction site are shown. The position of the Alu family repetitive sequence member is indicated by the box. The arrows below the restriction map indicate the position and orientation of short repeated sequences in this region.

```
        10         20         30         40         50         60
CTCGAGGGAG GAGCCCGGGG CTGGGGTACG GAGGCCTCTG CACATCTTAG AGTAAAACAA
GAGCTCCCTC CTCGGGCCCC GACCCCATGC CTCCGGAGAC GTGTAGAATC TCATTTTGTT
                                                   *
        70         80         90        100        110        120
GCAGGAGAGG CTGGGTGCGG TGGCTCATGC CTATAATCCC AGCACTTTAG GAGGCTGAGG
CGTCCTCTCC GACCCACGCC ACCGAGTACG GATATTAGGG TCGTGAAATC CTCCGACTCC

  *    130        140        150        160        170        180
CGGGCAGATC ACCTGAGGTC GGGAGTTCAA GACCAGCCTG ACCAACAGGG AGAAACCCCA
GCCCGTCTAG TGGACTCCAG CCCTCAAGTT CTGGTCGGAC TGGTTGTCCC TCTTTGGGGT

       190        200        210        220        230        240
TCTTTACTAA AACTACAAAA TTAGCTGGGT GTGGTGGCAC ATGCCTGTAA TCCCAGATAT
AGAAATGATT TTGATGTTTT AATCGACCCA CACCACCGTG TACGGACATT AGGGTCTATA

   *   250       *260        270        280        290        300
TCGGGAGGCT GAGGCAGGAG AATCGCTTGA ACCTGGGAAG CAGAGGTTGC GCTGAGCCGA
AGCCCTCCGA CTCCGTCCTC TTAGCGAACT TGGACCCTTC GTCTCCAACG CGACTCGGCT

       310        320        330        340        350        360
GATGGCACCA TTGCACTCCA GCCTGGGCAA CGAGAGCGAA ACTCCGTCTC AAAAAAACAA
CTACCGTGGT AACGTGAGGT CGGACCCGTT GCTCTCGCTT TGAGGCAGAG TTTTTTTGTT

       370        380        390        400        410        420
AAACAAAAAA ATCAAAACAA TCAAAAAAAC AAGCAGGAGG GGCTCTGAGG TGCCTGCAAC
TTTGTTTTTT TAGTTTTGTT AGTTTTTTTG TTCGTCCTCC CCGAGACTCC ACGGACGTTG

       430        440        450        460        470        480
ACCCAGGTAC AATCCGTGGC CCTGAGGCCC ATCACAGGGA AGGGGTCTTT GCAGCTCTTT
TGGGTCCATG TTAGGCACCG GGACTCCGGG TAGTGTCCCT TCCCCAGAAA CGTCGAGAAA

       490        500        510        520        530        540
CAACCCCCAG CCCAGCATCC AAGGAAGCCC AGGGCAGGGA GAAACCTCAG CTGCACCATC
GTTGGGGGTC GGGTCGTAGG TTCCTTCGGG TCCCGTCCCT CTTTGGAGTC GACGTGGTAG

       550        560        570        580        590        600
AGAGCTCAGA ACAGAGAAGG CAGAAATTAG CAGGGAGTGG GGCTGGGGAG GCTTCCTAGA
TCTCGAGTCT TGTCTCTTCC GTCTTTAATC GTCCCTCACC CCGACCCCTC CGAAGGATCT

       610        620        630        640        650        660
AGACGTGTCT CCCGCCTTGC TGGCACTGAG GCCTTGAGGA TGGGTCCATA CTGGGCCCCC
TCTGCACAGA GGGCGGAACG ACCGTGACTC CGGAACTCCT ACCCAGGTAT GACCCGGGGG

       670        680        690        700        710        720
ACTGCCAGGG ATGCAGATCC GGCCCACTGC TGAAATCTGT GCTCCTGGAG CCTCCCTCCT
TGACGGTCCC TACGTCTAGG CCGGGTGACG ACTTTAGACA CGAGGACCTC GGAGGGAGGA

       730        740        750        760        770        780
GTTCATGGGC CACAGGCTGT GAAAACCCCA GAGTCCTCCC AGGCAGCAAG TTTTGTTTTG
CAAGTACCCG GTGTCCGACA CTTTTGGGGT CTCAGGAGGG TCCGTCGTTC AAAACAAAAC

       790        800        810        820
TTTTTTGTTT GTTTGCTTGT TTGTTTTTTG AGAGTCTGCT CGTCA
AAAAAACAAA CAAACGAACA AACAAAAAAC TCTCAGACGA GCAGT
```

FIGURE 7. Sequence of a member of the Alu family of dispersed middle repetitive human DNA sequences and flanking regions. The 19 bp direct repeats at the boundaries of this member are indicated by double underlining. The 83 bp duplicated sequence is underlined. Asterisks mark the boundaries of the putative 14 bp replication origins (see text).

```
                Internal homology in human Alu family repeat unit

                   386 AA AACAAGCAGG AGGGGCT
                       ** ********** ** ****
 36 CTCTGCACAT CTTAGAGTAA AACAAGCAGG AGAGGCTGGG TGCGGTGGCT CATGCCTATA
         * *     *  * *** *** * **   *   ****** ** ******  ******* **
171 AGAAACCCCA TCTTTACTAA AACTA CAAA ATTAGCTGGG TGTGGTGGCA CATGCCTGTA

       SV40 ori CTCAGAGGC AGAGGCGGCC TC
                    *****  ******* *
 96 ATCCCAGCAC TTTAGGAGGC TGAGGCGGGC AGATCACCTG AGGTCGGGAG TTCAAGACCA
    *******    **  ****** ****** **    *** * ** *     ****      **
230 ATCCCAGATA TTCGGGAGGC TGAGGCAGGA GAATCGCTTG AACCTGGGAA GCAGAGGTTG

156 GCCTGACCAA CAGGGAGAAA CCCCATCTTT ACTAAAACTA CAAAAT
      **** *      *  *         **   *        *    *
290 CGCTGAGCCG AGATGGCACC ATTGCACTCC AGCCTGGGCA ACGAGA
%
```

FIGURE 8. Nucleotide sequence homologies within the Alu family repeat. The homology was maximized by inserting a space when necessary. Matching bases are indicated by asterisks. The position of each sequence segment in Fig. 7 is indicated by nucleotide number. Segments 386-404 and 54-72 are of the direct repeats which flank this Alu family member. Sequences 36-201 and 171-335 are the regions around and including the 83 bp internal duplication. The regions of homology with the SV40 origin of replication are also indicated.

hybridization to the adjacent 325 bp BglI fragment (Fig. 5, lane 2) although it contains 85 bp of the repeated sequence. This could reflect that sequence homology between members of the family is greatest in the region from position 73 to 300 and less from 300-385. If the size of the direct repeats at the boundary of this Alu family member is reduced to the 7 bp sequence, AAAACAA, there are 3 additional positions (Fig. 7, positions 354-360, 360-366 and 374-380) at which the right hand boundary of the direct repeat could be placed. Interestingly, the left boundary sequence (Fig. 7), positions 54-136, is homologous (62 of 83 nucleotides (75%)) to an internal sequence, positions 189-270 (Fig. 8). 12 of 19 nucleotides of the direct repeat flanking the Alu repeat are present in this internal sequence. These two homologous regions contain a 14 bp sequence (Fig. 7, positions 110-123 and 245-258) present near the origin of replication of several viruses including SV40 (Fig. 8) (20). The nucleotide sequence homology between members of the Alu family does not appear to be greater within this 83 bp duplicated region than in adjacent regions. A similar duplication is evident but the homology is lower in the partial sequence of another member of the Alu middle repetitive DNA family determined by Rubin et al. (11). The available data does indicate though that each Alu family member may contain two regions homologous to SV40 origin of replication.

The sequence (Fig. 7) also contains a 459 bp region (positions 351-809) whose boundaries are approximately 40 bp inverted repeats (positions 351-389 and 771-809). One boundary of this region lies within the Alu repeat unit. The relation of other Alu repeat members to structures like this is unknown, but all three members which have been sequenced contain an A-rich region which is a part of one of the two direct repeats (11,20). This region is reminiscent of several 5S and tRNA genes which are flanked on their 3' side by an dA-rich sequence in the coding strand and which is probably involved in transcription termination (21,22). We have examined the sequence within the inverted repeats for tRNA-like sequences and found none. As discussed above and indicated in Fig. 4A, lane 6, a portion of the sequence of 375 bp SacI is repeated elsewhere in the genome. We believe that the observed hybridization of this SacI fragment is of the essentially homopolymeric sequence, positions 771-809, which comprise one half of the inverted repeat. These data indicate that some of the inverted repeat (fold-back) structures observed in human DNA may be generated by complementary homopolymeric sequences (23).

Location of a Repetitive Sequence Flanking the 3' Side of the Rat Insulin I Gene

The position(s) of repeated sequence(s) in a 9.4 kb segment of rat DNA containing the insulin I gene (3) was determined by hybridization of labeled rat DNA to separated DNA fragments of the 9.4 kb segment (Fig. 9A,B). The data indicate that there is a repetitive sequence in a region 750-1850 bp from the 3' end of the gene (Fig. 6C). This rat repetitive sequence however does not hybridize with the human sequence (data not shown).

DISCUSSION

The sequences flanking the human insulin gene have been examined in order to assess their possible role in the regulation of insulin biosynthesis. We identified a region 700 bp from the 5' end into which there is apparent insertions of DNA. These insertions can vary from 100 bp to at least 2000 bp and in the one instance in which the inserted DNA has been analyzed, λHI-2 (Fig. 1), hybridization studies indicate that it is

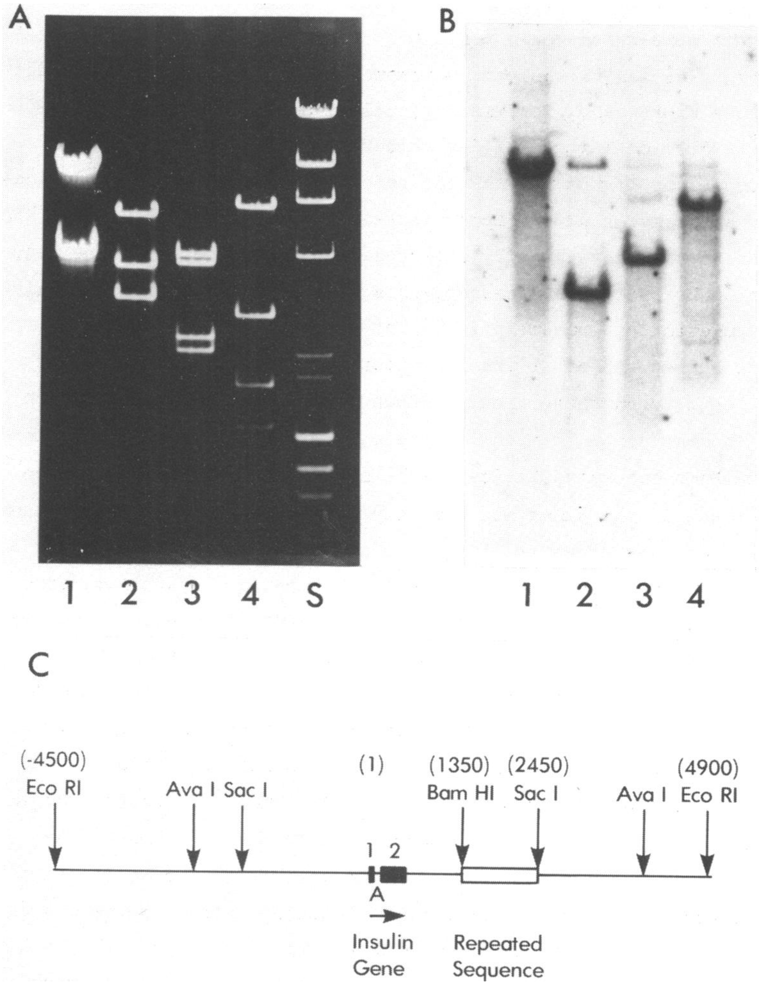

FIGURE 9. Identification of repeated sequences around the rat insulin I gene. The 9.4 kb EcoRI fragment containing the rat insulin I gene, subcloned in pBR322, was digested, electrophoresed in a 0.85% agarose gel, stained with ethidium bromide (A), transferred to nitrocellulose and hybridized with ^{32}P-labeled rat DNA (B). A restriction map of the fragment is shown in (C). The preproinsulin mRNA coding regions are indicated by solid boxes (1,2) and A is the intervening sequence. The arrow indicates the direction of transcription. The repeated sequence is located within the open box. The coordinates above the restriction map (in bp) are relative to the 5' end of the mRNA (position 1). The digests and fragment sizes (hybridizing fragments are underlined) are: 1. EcoRI-digested DNA, 9.4 kb and 4.4 kb (pBR322). 2. EcoRI-BamHI digested DNA, 5.9 kb, 4.0 kb (pBR322), and 3.5 kb. 3. EcoRI-SacI digested DNA, 4.4 kb (pBR322), 4.1 kb, 2.8 kb and 2.5 kb. 4. EcoRI-AvaI digested DNA, 6.4 kb, 2.9 (pBR322), 2.0 kb, 1.4 kb (pBR322) and 0.9 kb.

single copy sequence. The significance of insertions in the region flanking the gene which would presumably interact with transcriptional regulatory molecules is unclear, however they could conceivably affect insulin gene expression, for example, deletion of DNA greater than 2500 bp from its 5' side suppresses β-globin gene expression (24).

A comparison of the non-allelic rat insulin I and II genes indicated that they are homologous for approximately 500 bp on the 5' side of the genes whereas they have rapidly diverged on the 3' side (4). The sequence conservation before these genes suggests some common function possibly in the regulation of transcription. The analysis of heteroduplexes between the human and rat I insulin genes indicated that the sequence conservation between these insulin genes is restricted to the insulin encoding and immediate 5' flanking regions. These are also the regions which have been sequenced (1,3,4). The absence of extensive sequence homology flanking the human and rat insulin genes indicates that presumptive regulatory sequences are either conserved and close to the gene in the regions sequenced, that is within 100-150 bp, or less highly conserved and species specific. There is a similar absence of sequence conservation in the flanking regions of the chicken insulin gene as well (25).

We have identified the repeated sequences in a 19 kb region of chromosome 11 containing the 1430 bp insulin gene and including 5.6 kb before and 11.5 kb after the gene. Using two procedures of differing sensitivity, we observed that the repeated sequences are restricted to an approximately 500 bp region centered 6 kb from the 3' end of the insulin gene. There are no repeated sequences within 5.6 kb of the 5' end of the gene (this is the end of the cloned segment). We were also unable to detect sequence homology between subregions around the insulin gene, indicating the absence of localized, short range sequence repetition or pseudoinsulin gene sequences. The insulin gene then is embedded in a region of unique, nonrepetitive sequence of at least 11 kb.

The repetitive unit which is distal to the insulin gene possesses 79% sequence homology with clone 8 described by Rubin et al. (11) indicating that it belongs to the Alu family of dispersed middle repetitive sequences. This 313 bp sequence is bounded by 19 bp direct repeats. Jelinek et al. (20) partially characterized one of the seven members of the Alu family present in a 65 kb region of chromosome 11 which contains the β-globin gene cluster (12). The sequence of this member is homologous with the one described here but appears to terminate in 10 bp direct repeats whose sequence is different from the 19 bp repeats described here. This suggests that the mechanism by which the Alu family repeat is inserted into the chromosome is similar in that direct repeats are generated at each end of the segment but variable in that the length and sequence of the direct repeats are different. The dissimilarity in sequence of these short direct repeats indicates that they are probably not part of the Alu family member but are duplications of a region of the chromosome generated upon insertion. The determination of the sequences flanking other Alu family members will clarify this but it appears that Alu family members lack the terminal direct or inverted repeats which are part of transposable elements, for example the dispersed middle repetitive sequences Tyl in yeast (26) and copia, 412 and 297 in Drosophila (27,28), animal retroviruses (29), and bacterial insertion sequences and transposons (reviewed in Calos and Miller (30)). It is unknown whether members of the AluI family can also transpose, however if they do, they would be the smallest transposable elements which have been described and their mechanism of transposition must be different since they lack the terminal repetitive sequences characteristic of other transposing elements. We have examined two allelic insulin gene regions and both possess this repeated sequence in the same position. We have not determined its sequence in both alleles. This conservation of position of the Alu family repeat suggests that they are probably not highly mobile.

The function of the Alu family of middle repetitive sequences is unknown. The family is composed of approximately 300,000 members and comprises 3% of the human genome (9). At least some members, or portions of, are transcribed (20). This RNA appears restricted to the nucleus, and it has been suggested that it is involved in post-transcriptional processing of pre-messenger RNA. Also because of homology between a 14 bp segment of the Alu repeat and a region in the vicinity of the origin of replication of papovaviruses, Jelinek et al. (20) proposed that these middle repetitive sequences are origins of chromosome replication. The data presented here as well as that of Rubin et al. (11) indicates that there are two 14 bp segments in the Alu family with homology with viral origins of replication, not a single region as indicated by Jelinek et al. (20). Doolittle and Sapienza (31) and Orgel and Crick (32) have proposed that middle repetitive DNA sequences may be examples of selfish genes and may not possess a phenotype upon which natural selection can operate but may be preferred replicators. The Alu family of sequences has the properties proposed for selfish genes, i.e., multicopy sequences (genes) possessing origins of replication. In the 65 kb DNA segment containing the β-globin gene cluster, there are seven repeated sequences and all are probably Alu family members (12). They are located within the intergenic regions and appear to separate this segment of chromosome 11 into three regions: embryonic, fetal and adult β-globin genes. However it remains to be determined whether this organization has a functional role in the temporal regulation of β-globin gene expression. In the 19 kb segment of chromosome 11 around the insulin gene, we have identified only one member of the Alu family. The position of the next member is greater than 13.2 kb on one side and 5.5 kb on the other (on a statistical basis they would be expected to occur once every 6000 bp). We are presently trying to determine the positions of Alu family members at well as the location and identity of other genes around the human insulin gene to ascertain whether Alu family members are the boundaries of a DNA segment which contains genes expressed only in B cells, of the pancreas.

ACKNOWLEDGEMENTS

G.B. is the recipient of the Elliott P. Joslin Research and Development Award from the American Diabetes Association, Inc. We thank Leslie Spector for her assistance. This research was supported by a grant from Eli Lilly, Inc.

REFERENCES

1. Bell, G.I., Pictet, R.L., Rutter, W.J., Cordell, B., Tischer, E., & Goodman, H.M. (1980). Nature 284, 26-32.
2. Owerbach, D., Bell, G.I., Rutter, W.J. & Shows, T. (1980). Nature 286, 82-84.
3. Cordell, B., Bell, G., Tischer, E., DeNoto, F.M., Ullrich, A., Pictet, R., Rutter, W.J. & Goodman, H.M. (1979). Cell 18, 533-543.
4. Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R. & Tizard, R. (1979). Cell 18, 545-558.
5. Deininger, P.L. & Schmid, C.W. (1979). J. Mol. Biol. 127, 437-460.
6. Bostock, C.J., Gosden, J.R. & Mitchell, A.R. (1978). Nature 272, 324-328.
7. Manuelidis, L. (1978). Chromosoma 66, 1-21.
8. Mitchell, A.R., Beauchamp, R.S. & Bostock, C.J. (1979). J. Mol. Biol. 135, 127-149.

9. Houck, C.M., Rinehart, F.P. & Schmid, C.W. (1979). J. Mol. Biol. **132**, 289-306.
10. Cooke, H.J. & Hindley, J. (1979). Nucl. Acids Res. **6**, 3177-3197.
11. Rubin, C.M., Houck, C.M., Deininger, P.L., Friedmann, T. & Schmid, C.W. (1980). Nature **284**, 372-374.
12. Fritsch, E.F., Lawn, R.M. & Maniatis, T. (1980). Cell **19**, 959-972.
13. Rigby, P.W.J., Dieckmann, M., Rhodes, C. & Berg, P. (1977). J. Mol. Biol. **113**, 237-251.
14. Alwine, J.C., Kemp, D.J., Parker, B.A., Reiser, J., Renart, J., Stark, G.R. & Wahl, G.M. (1979). Methods in Enzymol. **68**, 220-242.
15. Southern, E.M. (1975). J. Mol. Biol. **98**, 503-517.
16. Wahl, G.M., Stern, M. & Stark, G.R. (1979). Proc. Natl. Acad. Sci. USA **76**, 3683-3687.
17. Maxam, A.M. & Gilbert, W. (1977). Proc. Natl. Acad. Sci. USA **74**, 560-564.
18. Lauer, J., Shen, C.J. & Maniatis, T. (1980). Cell **20**, 119-130.
19. Shen, C.J. & Maniatis, T. (1980). Cell **19**, 379-391.
20. Jelinek, W.R., Toomey, T.P., Leinwand, L., Duncan, C.H., Biro, P.A., Choudary, P.V., Weissman, S.M., Rubin, C.M., Houck, C.M., Deininger, P.L. & Schmid, C.W. (1980). Proc. Natl. Acad. Sci. USA **77**, 1398-1402.
21. Valenzuela, P., Bell, G.I., Masiarz, F.R., DeGennaro, L.J. & Rutter, W.J. (1977). Nature **267**, 641-643.
22. Valenzuela, P., Venegas, A., Weinberg, F., Bishop, R. & Rutter, W.J. (1978). Proc. Natl. Acad. Sci. USA **75**, 190-194.
23. Deininger, P.L. & Schmid, C.W. (1976). J. Mol. Biol. **106**, 773-790.
24. Van der Ploeg, L.H.T., Konings, A., Oort, M., Roos, D., Bernini, L. & Flavell, R.A. (1980). Nature **283**, 637-642.
25. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980). Cell **20**, 555-566.
26. Cameron, J.R., Loh, E.Y. & Davis, R.W. (1979). Cell **16**, 739-751.
27. Finnegan, D.J., Rubin, G.M., Young, M.W. & Hogness, D.S. (1978). Cold Spring Harbor Symp. Quant. Biol. **42**, 1053-1063.
28. Potter, S.S., Brorein, W.J., Dunsmuir, P. & Rubin, G.M. (1979). Cell **17**, 415-427.
29. Shimotohno, K., Mizutani, S. & Temin, H.M. (1980). Nature **285**, 550-554.
30. Calos, M.P. & Miller, J.H. (1980). Cell **20**, 579-595.
31. Doolittle, W.F. & Sapienza, C. (1980). Nature **284**, 601-603.
32. Orgel, L.E. & Crick, F.H.C. (1980). Nature **284**, 604-607.