The EMBO Journal vol.4 no.1 pp.77–84, 1985

Architecture and anatomy of the chromosomal locus in human chromosome 21 encoding the Cu/Zn superoxide dismutase

Ditsa Levanon, Judy Lieman-Hurwitz, Naomi Dafni, Meir Wigderson, Levana Sherman, Yael Bernstein, Zehava Laver-Rudich, Efrat Danciger, Orna Stein and Yoram Groner

Department of Virology, The Weizmann Institute of Science, Rehovot, Israel

Communicated by Y. Groner

The SOD-1 gene on chromosome 21 and ~100 kb of chromosomal DNA from the 21q22 region have been isolated and characterized. The gene which is present as a single copy per haploid genome spans 11 kb of chromosomal DNA. Heteroduplex analysis and DNA sequencing reveals five rather small exons and four introns that interrupt the coding region. The donor sequence at the first intron contains an unusual variant dinucleotide 5′-G-C, rather than the highly conserved 5′-GT. The unusual splice junction is functional *in vivo* since it was detected in both alleles of the SOD-1 gene, which were defined by differences in the length of restriction endonuclease fragments (RFLPs) that hybridize to the cDNA probe. Genomic blots of human DNA isolated from cells trisomic for chromosome 21 (Down's syndrome patients) show the normal pattern of bands. At the 5′ end of gene there are the 'TATA' and 'CAT' promoter sequences as well as four copies of the -GGCGGG- hexanucleotide. Two of these -GC- elements are contained within a 13 nucleotide inverted repeat that could form a stem-loop structure with stability of −33 kcal. The 3′-non coding region of the gene contains five short open reading-frames starting with ATG and terminating with stop codons.

Key words: Down's syndrome/two SOD-1 alleles/RFLPs/ unusual splice junction/-GGCGGG-transcriptional signals

Introduction

Superoxide dismutase (SOD: EC 1.15.1.1) is present in most aerobic organisms. It catalyses the dismutation of superoxide: $O_2^- + O_2^- + 2H^+ \rightarrow H_2O_2 + O_2$ (reviewed by Fridovich 1978, 1979). Eukaryotic cells contain two distinct forms of SOD – a mitochondrial manganese-containing enzyme and a cytoplasmic copper/zinc-containing enzyme (SOD-1). The human SOD-1 is a dimer of 32 kd composed of two identical non-covalently linked subunits (Briggs and Fee, 1978) with known amino acid sequences (Jabusch *et al.*, 1980; Barra *et al.*, 1980). The gene locus for human SOD-1 was assigned to chromosome 21 (Tan *et al.*, 1973). This chromosome is involved in the most common genetic disease, known as trisomy 21 or Down's syndrome (Lejeune *et al.*, 1959). About one in 1000 newborn babies carries an extra copy of chromosome 21 and thereby suffers from Down's syndrome. The clinical symptoms of Down's syndrome are severe mental retardation, slow physical development, increased incidence of leukemia, high susceptibility to infections and some signs of premature aging (Martin 1978; Burgio *et al.*, 1981). In addition, almost all Down's syndrome patients develop by the age of 40 progressive dementia with symptoms and neurological abnormalities identical to Alzheimer's disease (Solitaire and Lamarche, 1966; Owens *et al.*, 1971; Burger and Vogel, 1973; Ellis *et al.*, 1974; Heston, 1977), suggesting an association between the two conditions, i.e., genes on chromosome 21 could play a role in the development of Alzheimers disease. In most cases, the patients with Down's syndrome have a karyotype with 47 chromosomes (46 plus one additional 21). However, cases have been identified in which only a portion of chromosome 21 is present in triplicate, usually translocated to another chromosome. Those studies enabled the localization of the region 'responsible' for the syndrome on the chromosomal segment 21q22 or more specifically on q22.1 and possibly q22.2 (Niebuhr, 1974; Williams *et al.*, 1975; Cervenka *et al.*, 1977; Pfeiffer *et al.*, 1977; Poissonier *et al.*, 1976; Sinet *et al.*, 1976; Hagemeijer and Smith, 1977; Philip *et al.*, 1978; Tsukino *et al.*, 1980; Summit, 1981). Although the relationship of trisomy 21 to Down's syndrome has been known for 25 years (Lejeune *et al.*, 1959), there is no effective treatment and very little is known about the way in which the additional chromosomal segment (21q22) causes the disease. It is generally assumed that the extra 21q22 segment codes for normal products and that the abnormalities found in Down's syndrome are produced by the excess of some of those proteins. Indeed, the patients show an increase of ~50% in SOD-1 activity due to higher levels of SOD-1 protein (Sinet *et al.*, 1974; Eriksson *et al.*, 1975; Crosti *et al.*, 1976; Feaster *et al.*, 1977). Molecular cloning of the genes residing in the 21q22 region and analysis of their organization and expression should lead to the identification of those particular genes involved in the pathology associated with the syndrome. Because the gene locus of SOD-1 is located in the 21q22 segment this gene can serve as a convenient starting point for such an endeavor. In addition, since it was suggested that the overproduction of SOD-1 may be involved in some of the clinical symptoms of Down's syndrome (Sinet, 1982), we envisaged that cloning and characterization of the gene locus coding the SOD-1 may contribute to our understanding of this genetic disease. To this end, we have constructed a cDNA clone of human SOD-1 and studied the expression of the SOD-1 gene in different cells (Lieman-Hurwitz *et al.*, 1982; Sherman *et al.*, 1983). Two SOD-1 mRNAs of ~0.7 kb and 0.9 kb differing in the length of their 3′-untranslated region were found in a variety of human cells. They are transcribed from the same gene and the major 0.7 kb species is approximately four times more abundant than the 0.9-kb mRNA. Here we report the isolation and characterization of the SOD-1 gene locus including ~100 kb from the 21q22.1 chromosomal segment.

Results

Organization of the SOD-1 gene

Human DNA from FS11 cells or placenta was digested with *Eco*RI, *Bgl*II, *Hind*III or *Pst*I, fractionated by gel electrophoresis and transferred to nitrocellulose filters. When the

© IRL Press Limited, Oxford, England.

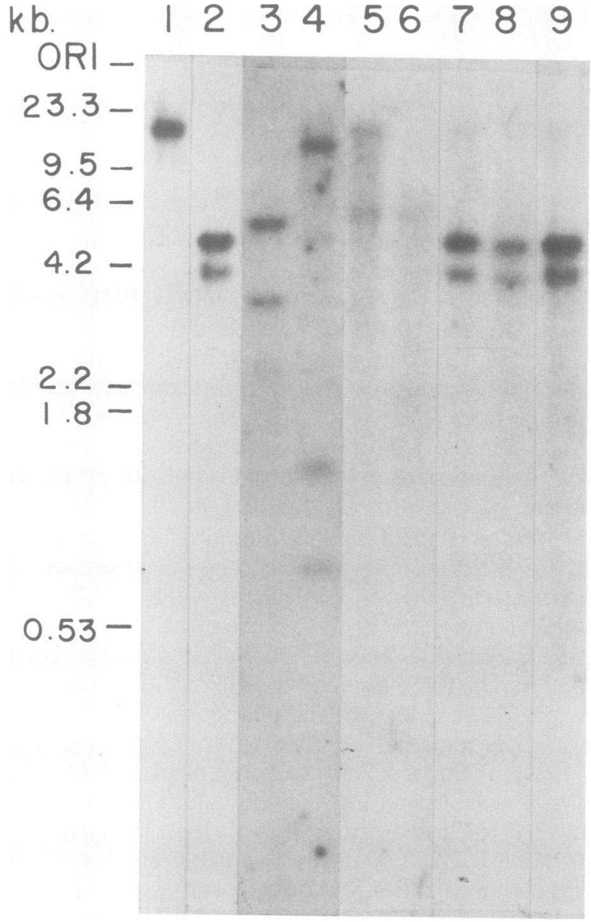

Fig. 1. Hybridization of ^{32}P-labeled SOD-1 cDNA to human genomic DNA. 20 μg of DNA were digested with various restriction enzymes, the fragments separated on a 0.8% agarose gel, transferred to nitrocellulose and hybridized to a ^{32}P-labeled probe of the SOD-1 cDNA (insert of pS61-10). **Lanes 1–4**, DNA was isolated from cultured FS-11 normal human fibroblasts and digested with **(1)** *Eco*RI; **(2)** *Bgl*II; **(3)** *Pst*I; **(4)***Hind*III. **Lanes 5** and **6**, DNAs were digested with *Eco*RI; **(5)** WAVR4dF9-4a mouse-human hybrid cells (Kozak *et al.*, 1977) **(6)** A9, mouse cell. **Lanes 7–9**, DNAs were digested with *Bgl*II; **(7)** SV-80 cells; **(8)** WAV4dF9-4a **(9)** human placenta.

blot was hybridized with ^{32}P-labeled SOD-1 cDNA (pS61-10) (Lieman-Hurwitz *et al.*, 1982; Sherman *et al.*, 1983) hybridizations were observed to fragments of 18 kb-(*Eco*RI); 5.1 and 4.1 kb-(*Bgl*II); 5.6, 3.6 and 2.3 kb-(*Pst*I); 14.5, 1.3 and 0.8 kb-(*Hind*III) (Figure 1, lanes 1,2,3, and 4). Hybridization under less stringent conditions revealed additional bands (hardly visible in Figure 1, but see Figure 7). These fragments were derived from SOD-1 related processed genes and possess only partial homology to the SOD-1 sequences (Danciger *et al.*, in preparation). All the SOD-1 related DNA fragments were isolated as recombinant λ phages by screening ~1 x 10^6 phage plaques from each of two genomic libraries: one was prepared from partial *Hae*III and *Alu*I digests of human fetal liver DNA by Lawn *et al.* (1978) and the other was constructed from a partial *Eco*RI digest (Mory *et al.*, 1981). Eleven separated Charon 4A phage clones containing SOD-1 related sequences were isolated. Detailed restriction mapping and hybridization to the cDNA as well as heteroduplex analyses indicated that four overlapping phage clones designated λA-2, λB-1, λF and λ5-1 originated from the SOD-1 chromosomal locus (Figure 2). This region was enlarged by several rounds of 'chromosome walking' (see Hadfield 1983) using both libraries so that a total of ~100 kb of chromosomal DNA from the 21q22.1 segment were isolated (Figure 2). Experiments aimed to detect the SOD-1 neighboring gene by hybridizing Northern blots with ^{32}P-labeled genomic fragments have yielded, so far, negative results. The SOD-1 gene is contained within one large *Eco*RI fragment of ~18 kb (λA-2). Obviously this phage and hence the SOD-1 gene is underrepresented in any partial *Eco*RI phage library. The 18-kb *Eco*RI fragment containing the SOD-1 gene was also detected in DNA from the mouse/human hybrid line WAVR4dF9-4a that contains human chromosome 21 as the only human chromosome (Figure 1, lane 5). The additional 6-kb fragment seen in this lane co-migrated with the mouse fragment (lane 6). When DNA from human trisomy-21 cells was digested with *Eco*RI, *Bgl*II or *Pst*I, fractionated on agarose gel and analyzed by blot hybridization, the pattern obtained was qualitatively identical to that of the disomy fibroblasts depicted in Figure 1, indicating that there are no gross differences associated with the extra copy of the SOD-1 gene.

The organization of the SOD-1 gene was examined by electron microscopy of heteroduplexes formed between a fragment isolated from the recombinant phage λB-1 and the SOD-1 cDNA clone. A 13-kb *Eco*RI-*Eco*RI fragment containing the entire SOD-1 gene was isolated from λB-1 and subcloned in pHG165 (a pBR322 derivative containing a poly-linker provided by J. Kuhn). It was then linearized by *Bam*HI which also trimmed a 1.5-kb fragment of the 3′-flanking region beyond the SOD-1 gene, hybridized with the linearized SOD-1 cDNA subcloned in pHG165 and examined by electron microscopy. Five blocks of homologous sequences and four intron loops were visualized on the electron micrographs (Figure 3). The 5′-3′ orientation of the gene shown in Figure 3 was deduced from the known orientation of the cDNA clone (Lieman-Hurwitz *et al.*, 1982; Sherman *et al.*, 1983). The double-stranded tail of the vector pHG165 (3.4 kb) is followed by loop A (0.5 kb) which defines the distance between the 5′ terminus of the genomic clone and the first exon. Loops B through E (totalling 3.1 kb) represent the first intron. The stem-loop structures C and D are formed by two pairs of inverted repeats within the first intron. Loops F (2.0 kb), H (0.45 kb), and G (1.8 kb) represent the second, third and fourth introns, whereas the single-stranded tail at the 3′ side represents the 3′-flanking region of the gene.

Nucleotide sequence of the SOD-1 gene

To locate and isolate smaller DNA fragments containing the SOD-1 sequences, cloned DNAs (λA-2, λB-1, λF and λ5-1) were digested with a variety of restriction enzymes and analysed by blot hybridization with ^{32}P-labeled SOD-1 cDNA. The relevant regions were subcloned in pBR322 or pUC13 and the exons were more precisely localized on the genomic DNA (Figure 2). To establish the exact exon-intron structure of the SOD-1 gene the nucleotide sequence of all the exons and part of their flanking introns were determined using both the chemical method of Maxam and Gilbert (1980) and the M13 phage dideoxynucleotide technique (Sanger *et al.*, 1980) (Figure 4). As predicted by the heteroduplex analyses the coding region of the human SOD-1 gene is divided into five exons interrupted by four introns. The nucleotide sequence of the five genomic exons is identical to that of the previously published SOD-1 cDNA (pS61-10) (Sherman *et al.*, 1983) indicating that the cloned DNAs (λA-2 and λB-1)

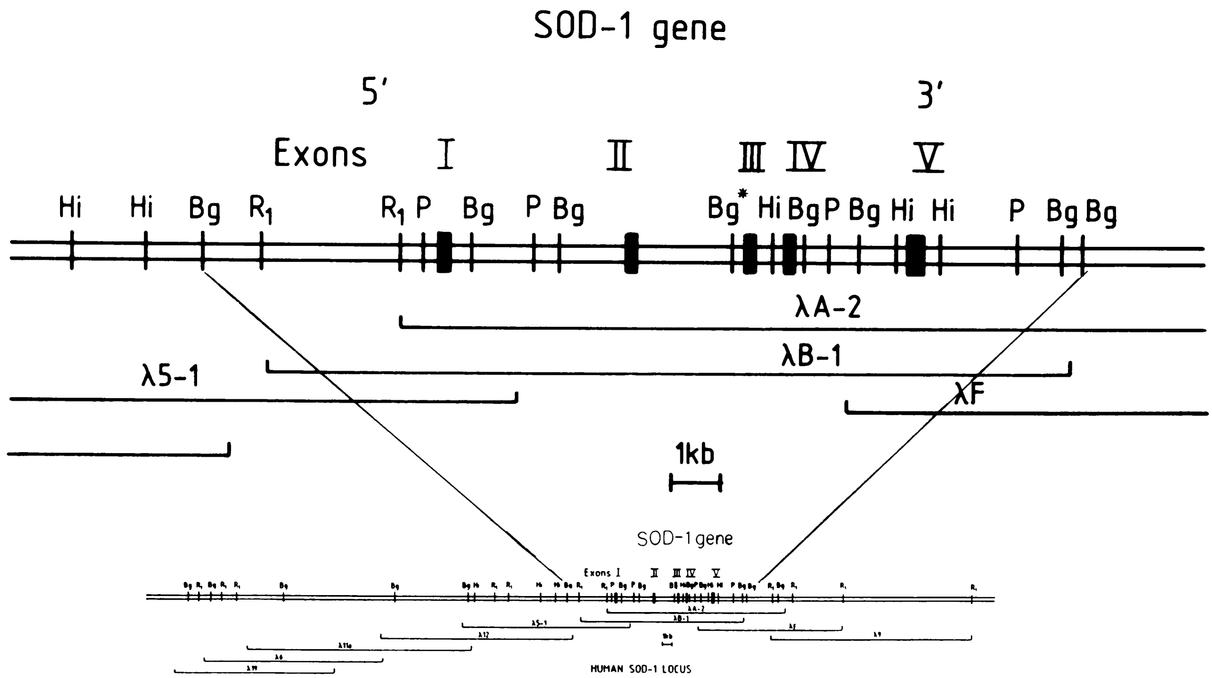

Fig. 2. Restriction map and organization of the SOD-1 gene locus. Top line: genomic DNA from four overlapping λ clones with positions of four restriction enzymes R_1 = *Eco*RI, Bg = *Bgl*II, Hi = *Hind*III, P = *Pst*I. The exons are the black rectangles numbered I – V. The polymorphism in the *Bgl*II site is indicated by an asterisk. Bottom line: molecular map of ~100 kb of the SOD-1 locus. The nine overlapping λ clones contain sequences of the gene and neighboring regions.

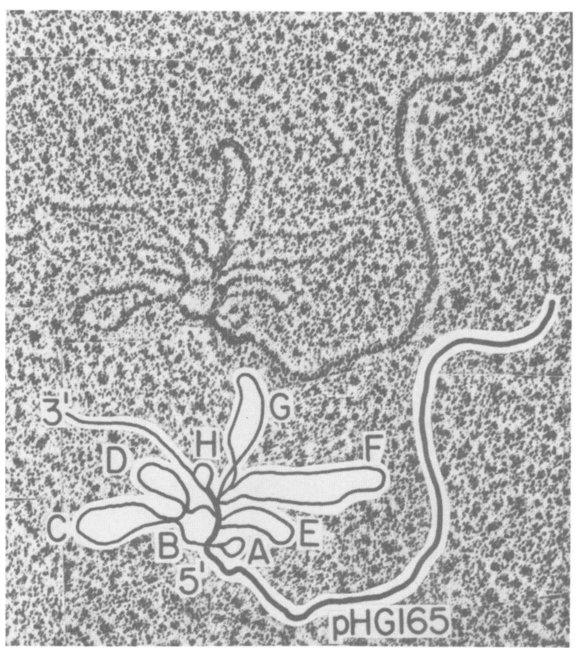

Fig. 3. Electron micrograph and tracing of heteroduplex between the SOD-1 genomic and cDNA clones. The two recombinant clones, both subcloned in pHG165, were denatured, annealed and mounted for microscopy. Single- and double-stranded DNAs are represented by thin and thick solid lines, respectively. See the text for the (A) to (H) loops and 5′-3′ orientation.

represent the functional gene. In Figure 5 the sequences of the donor and acceptor splice junctions at the exon-intron boundaries are compared with the proposed consensus sequence for such junctions (Breathnach and Chambon, 1981; Mount, 1982). The donor sequence at the first intron contains an unusual variant dinucleotide 5′-G-C rather than the highly conserved GT; whereas all the other three introns are bound by the consensus 5′-G-T. . . . . . A-G-3′ (Figure 5). The sequencing of the region containing the unusual 5′-G-C donor site was repeated several times from both strands; the results were identical in all experiments. Among the other nucleotides the 5′-A-G preceding the invariant -G-T of the donor sequence appears in three out of the four introns and the boundaries of intron no.4 are in very good agreement with the consensus sequence. The unusual 5′-G-C dinucleotide in the first intron is not an artifact of the cloning procedure because it was detected in both λA-2 and λB-1 which are alleles of the SOD-1 gene (see below).

The 5′- and 3′-terminal regions of the SOD-1 gene

Exon 1 contains 72 nucleotides of the coding region and 80 nucleotides corresponding to the mRNA 5′-untranslated region. The mRNA start site indicated in Figure 4 was determined by both S1 nuclease mapping and primer extension experiments (data not shown). The genomic sequence upstream from the mRNA start site contains two conserved sequences involved in the promotion of transcription by RNA polymerase II (Breathnach and Chambon, 1981). At 23 – 29 nucleotides upstream from the mRNA start site there is a hexanucleotide T-A-T-A-A-A known as the 'Goldberg-Hogness' box and at positions −69 and −128 relative to the mRNA start site there are C-C-A-T-T and C-A-T-T sequences, respectively (Figure 4). In addition to the 'TATA' and 'CAT' boxes, the promoter region of the SOD-1 gene contains three direct and one inverted repeats of the sequence 5′-GGCGGG-3′ at nucleotide nos. −90, −135, −172 and −243 (underlined in Figure 4). Also present in this region are two pairs of inverted repeats: a shorter one of seven nucleotides between the 'TATA' and 'CAT' boxes (marked by two arrows in Figure 6) and a longer 13 nucleotide pair preceding the −69 'CAT' box. Interestingly, the −90 'GC' and −135

```
GTACCCTGTT TACATCATTT TGCCATTTTC GCGTACTGCA ACCGGCGGGC CACGCCGTGA AAAGAAGGTT GTTTTCTCCA CAGTTTCGGG GTTCTGGACG TTTCCCGGCT

GCGGGGCGGG GGGAGTCTCC GGCGCACGCG GCCCCTTGGC CCGCCCCAGT CATTCCCGGC CACTCGCGAC CCGAGGCTGC CGCAGGGGGC GGGCTGAGCG CGTGCGAGGC
                                                                                          mRNA start site
CATTGGTTTG GGGCCAGAGT GGGCGAGGCG CGGAGGTCTG GCCTATAAAG TAGTCGCGGA GACGGGGTGC TGGTTTGCGT CGTAGTCTCC TGCAGGTCTG GGGTTTCCGT

                                                    Ala Thr Lys Ala Val Cys Val Leu Lys Gly Asp Gly Pro Val Gln Gly Ile
TGCAGTCCTC GGAACCAGGA CCTCGGCGTG GCCTAGCGAG TT ATG GCG ACG AAG GCC GTG TGC GTG CTG AAG GGC GAC GGC CCA GTG CAG GGC ATC

Ile Asn Phe Glu Gln Lys
ATC AAT TTC GAG CAG AAG GCAAGGGCTG GGACCGGGAG GCTTGTGTTG CGAGGCCGCT CCCGACCCGC TCGTCCCCCC GCGACCCTTT GCATGGACGG GTCGCCCGCC

AGGG.........................CCTAGAGCAGT TAAGCAGCTT GCTGGAGGTT CACTGGCTAG AAAGTGGTCA GCCTGGGATT TCGGACACAG ATTTTTCCAC

                                                                                                      Glu Ser Asn
TCCCAAGTCT GGCTGCTTTT TACTTCACTG TGAGGGGTAA AGGTAAATCA GCTGTTTTCT TTGTTCAGAA ACTCTCTCCA ACTTTGCACT TTTCTTAAAG GAA AGT AAT

Gly Pro Val Lys Val Trp Gly Ser Ile Lys Gly Leu Thr Glu Gly Leu His Gly Phe His Val His Glu Phe Gly Asp Asn Thr Ala
GGA CCA GTG AAG GTG TGG GGA AGC ATT AAA GGA CTG ACT GAA GGC CTG CAT GGA TTC CAT GTT CAT GAG TTT GGA GAT AAT ACA GCA G GTGG

                                                            Gly Cys Thr Ser Ala Gly Pro His Phe Asn Pro Leu
GT..............................CATAATTTAG CTTTTTTTTC TTCTTCTTAT AAATAG GC TGT ACC AGT GCA GGT CCT CAC TTT AAT CCT CTA

Ser Arg Lys His Gly Gly Pro Lys Asp Glu Glu Arg
TCC AGA AAA CAC GGT GGG CCA AAG GAT GAA GAG AG GTAACAAGAT GCTTAACTCT TGTAATCAAT GGCGATACGT TTCTGGAGTT CATATGGTAT ACTACTTGTA

AATATGTGCC TAAGATAATT CCGTGTTTCC CCCACCTTTG CTTTTGAACT TGCTGACTCA TGTGAAACCC TGCTCCCAAA TGCTGGAATG CTTTTACTTC CTGGGCTTAA

AGGAATTGAC AAATGGGCAC TTAAAACGAT TTGGTTTTGT AGCATTTGAT TGAATATAGA ACTAATACAA GTGCCAAAGG GGAACTAATA CAGGAAATGT TCATGAACAG

                                                                                        His Val Gly Asp Leu Gly
TACTGTCAAC CACTAGCAAA ATCAATCATC ATT.................................TGATGCTTTT CATATAG G CAT GTT GGA GAC TTG GGC

Asn Val Thr Ala Asp Lys Asp Gly Val Ala Asp Val Ser Ile Glu Asp Ser Val Ile Ser Leu Ser Gly Asp His Cys Ile Ile Gly Arg
AAT GTG ACT GCT GAC AAA GAT GGT GTG GCC GAT GTG TCT ATT GAA GAT TCT GTG ATC TCA CTC TCA GGA GAC CAT TGC ATC ATT GGC CGC

Thr Leu Val
ACA CTG GTG GTAAGTTTTC ATAAAGGATA TGCATAAAAC TTCTTCTAAC AGTACAGTCA TGTATCTTTC ACTTTGATTG TTAGTCGCGA ATTCTAAGAT CCAGATAAAC

TGT............................................GTTTCTGCTT TTAAACTACT AAATATTAGT ATATCTCTCT ACTAGGATTA ATGTTATTTT

                                                                                                      Val His Glu
TCTAATATTA TGAGGTTCTT AAACATCTTT TGGGTATTGT TGGGAGGAGG TAGTGATTAC TTGACAGCCC AAAGTTATCT TCTTAAAATT TTTTACAG GTC CAT GAA

Lys Ala Asp Asp Leu Gly Lys Gly Gly Asn Glu Glu Ser Thr Lys Thr Gly Asn Ala Gly Ser Arg Leu Ala Cys Gly Val Ile Gly Ile
AAA GCA GAT GAC TTG GGC AAA GGT GGA AAT GAA GAA AGT ACA AAG ACA GGA AAC GCT GGA AGT CGT TTG GCT TGT GGT GTA ATT GGG ATC

Ala Gln STOP                                                          I
GCC CAA TAAACATTCC CTTGGATGTA GTCTGAGGCC CCTTAACTCA TCTGTTATCC TGCTAGCTGT AGAAATGTAT CCTGATAAAC ATTAAACACT GTAATCTTAA

Poly A                                                               II                               III
AAGTGTAATT GTGTGACTTT TTCAGAGTTG CTTTAAAGTA CCTGTAGTGA GAAACTGATT TATGATCACT TGGAAGATTT GTATAGTTTT ATAAAACTCA GTTAAAATGT

                                                                                                 IV   Poly A
CTGTTTCAAT GACCTGTATT TTGCCAGACT TAAATCACAG ATGGGTATTA AACTTGTCAG AATTTCTTTG TCATTCAAGC CTGTGAATAA AAACCCTGTA TGGCACTTAT

TATGAGGCTA TTAAAAGAAT CCAAATTCAA ACTAAATTAG CTCTGATACT TATTTATATA AACAGCTTCA GTGGAACAGA TTTAGTAATA CTAACAGTGA TAGCATTTTA

                                 V
TTTTGAAAGT GTTTTGAGAC CATCAAAATG CATACTTTAAAACAGCAGGTC TTTTAGCTAA AACTAACACA ACTCTGCTTA GACAAATAGG CTGTCCTTTG  AAGCTT
```

Fig. 4. Nucleotide and amino acid sequences of the SOD-1 gene. The sequence of all coding regions and adjacent nucleotides are shown with 110 bases per line. The 'TATA', 'CAT' and polyadenylation sequences are boxed. The splice junctions are underlined. The exons were identified by comparison with the cDNA sequences (Sherman *et al.*, 1983). The sites of initiation of transcription and poly(A) are indicated. The arrows mark the two 9-nucleotide direct repeats at the 3′-non-coding region.

'CG' elements are contained in the paired region of a putative stem-loop structure that could be formed by the 13 nucleotides repeats, whereas, the −128 'CAT' box is located in the unpaired loop (Figure 6). The stability of this structure was estimated according to Tinoco *et al.* (1973) as −33 kcal. The 7-bp repeat on the other hand could form a less stable structure of −17 kcal. Such 'GC' elements and stem-loop structures could serve as a binding or recognition site for regulatory proteins and thus may play a role in the regulation of SOD-1 expression. The fifth exon includes the last 35 codons followed by a TAA termination triplet and 312 bp of the 3′-untranslated region (Figure 4). Downstream from the stop codon there are 76 nucleotides before the hexanucleotide ATTAAA. This polyadenylation signal is located 21 bp upstream from the poly(A) tail identified by Sherman *et al.* (1983) in the SOD-1 cDNA clone and is the one involved in the formation of the 0.7-kb SOD-1 mRNA species. There are three additional poly(A) signals located 200 – 250 bp further downstream (Figure 4). The middle one (at nucleotide no. 1824) specifies the poly(A) site of the 0.9-kb SOD-1 mRNA (Sherman *et al.*, in preparation). Curiously, the 3′-untranslated region of the SOD-1 is marked by two unique features: first, immediately at the end of the coding region there is a 67-bp fragment flanked by two perfect 9-nucleotides direct repeats indicated by the two arrows in Figure 4; second, five short open reading frames starting with ATG (boxed and numbered I – V) and terminating with stop codons were identified beyond the end of the SOD-1 coding sequence. The shortest reading-frame (no.1) contains six codons, while the longest one (no.3) consists of 29 codons.

The phage clones λA-2 and λB-1 are alleles of the SOD-1 gene

As indicated above, the λA-2 and λB-1 recombinants were isolated from the library of human fetal liver DNA (Lawn *et al.*, 1978) and both contained the entire SOD-1 gene (Figure 2). When hybridized to each other these two recombinants formed a stable DNA duplex across their overlapping region indicated in Figure 2. The restriction maps of that region are identical except for one *Bgl*II site (marked by the asterisk in Figure 2) which is missing in λB-1. When the two phage DNAs were cut by *Bgl*II, blotted and probed with the SOD-1 cDNA clone the λA-2 generated three fragments of 4.1 kb, 3.6 kb and 1.5 kb. The 3.6-kb and 1.5-kb *Bgl*II fragments were missing from the digest of λB-1, instead it contained one 5.1-kb fragment (Figure 2). In the genomic blot shown in Figure 1, only the 5.1-kb and 4.1-kb fragments are visible in the various human DNAs (Figure 1, lanes 2, 7, 8 and 9), indicating that these samples contain the λB-1 form of the SOD-1 gene. To test whether the λA-2 and λB-1 are alleles of the same locus, seven different human DNA samples from unrelated individuals were digested with *Bgl*II and analyzed by Southern blot hybridization. Two out of the seven samples contained the 1.5-kb fragment which is diagnostic for the λA-2 form of the gene (Figure 7). We, threfore, concluded that the λA-2 and λB-1 are alleles and that the additional *Bgl*II site in λA-2 was created by an alteration in the nucleotide sequence which causes the well-documented restriction fragment length polymorphism (RFLPs) (Kan and Dozy, 1978; Botstein *et al.*, 1980). The Mendelian inheritance of this RFLP, as well as other polymorphic DNA markers present in the SOD-1 locus, were determined by analyzing the segregation patterns in informative families (Antonarakis *et al.*, unpublished results). The two other fragments present in all the DNA samples in Figure 7, i.e., the large 17-kb and the 3.6-kb bands, belong to the SOD-1 related pseudogenes mentioned above. They are clearly visible here due to the lower stringency of the hybridization conditions. These two SOD-1 related pseudogenes, as well as two additional pseudogenes, have been isolated from the human λ phage library and characterized (Danciger *et al.*, in preparation).

	DONOR(GT) 5′ 3′	ACCEPTOR(AG) 5′ 3′
CONSENSUS	(A/C)AGGT(A/G)AGT	(C/T)(C/T)(C/T)(C/T)(C/T)(C/T)X(C/T)AGG(G/T)
INTRON #1	AAG**GC**AAGG	TTTCTTAAAGGA
" #2	CAGGTGGGT	TTATAAATAGGC
" #3	GAGGTAACA	TTTCATATAGGC
" #4	GTGGTAAGT	TTTTTTACAGGT

Fig. 5. Exon-intron junctions of the SOD-1 gene. The nucleotide sequences bordering the coding regions of the SOD-1 gene are compared with the consensus sequence. The unusual 'GC' variant in the donor of the first intron is marked by bolder letters.

Discussion

Human genomic libraries in lambda Ch4A were screened with cloned SOD-1 cDNA and a few overlapping recombinant phages containing the whole SOD-1 gene were isolated. The DNA regions present in λA-2 and λB-1 are the only ones among the phages that were picked-up with nucleotide sequences identical to the SOD-1 cDNA. We therefore concluded that this region represents the unique SOD-1 functional gene. The gene is ~11 kb in length and is interrupted by four introns. In proportion to the sizes of the two SOD-1 mRNAs (0.7 and 0.9 kb) this is a large gene because it is over 12 times the length of the longer mRNA species. In addition to the functional gene we have isolated four processed pseudogenes. Experiments with genomic library of human chromosome 21 (Krumlauf *et al.*, 1982) have indicated that the processed genes do not reside on this chromosome (in preparation). Genomic blots of human DNAs isolated from cells trisomic for chromosome 21 show the normal pattern of bands after digestion with *Eco*RI, *Pst*I or *Bgl*II and hybridization to the SOD-1 cDNA probe. This is expected since it is assumed that the additional chromosome 21 codes for the normal cellular proteins and that the abnormalities observed in Down's syndrome are due to an excess of some of these gene products. All the protein coding regions and part of the introns of the

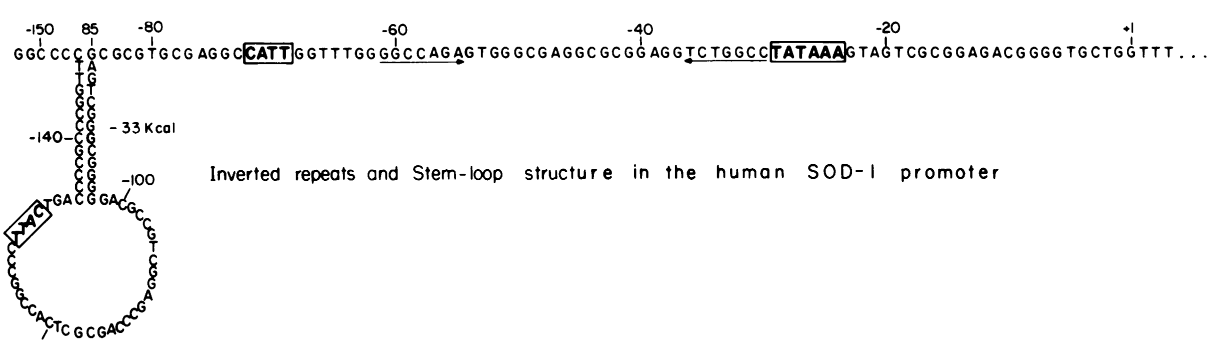

Fig. 6. Putative stem-loop structure located 5′ to the 'TATA' and 'CAT' boxes of the SOD-1 gene. The numbers correspond to those in Figure 4. 'TATA' and 'CAT' sequences are boxed. The free energy per strand (– 33 kcal/mol) is indicated. The two arrows mark the 7-bp inverted repeat.

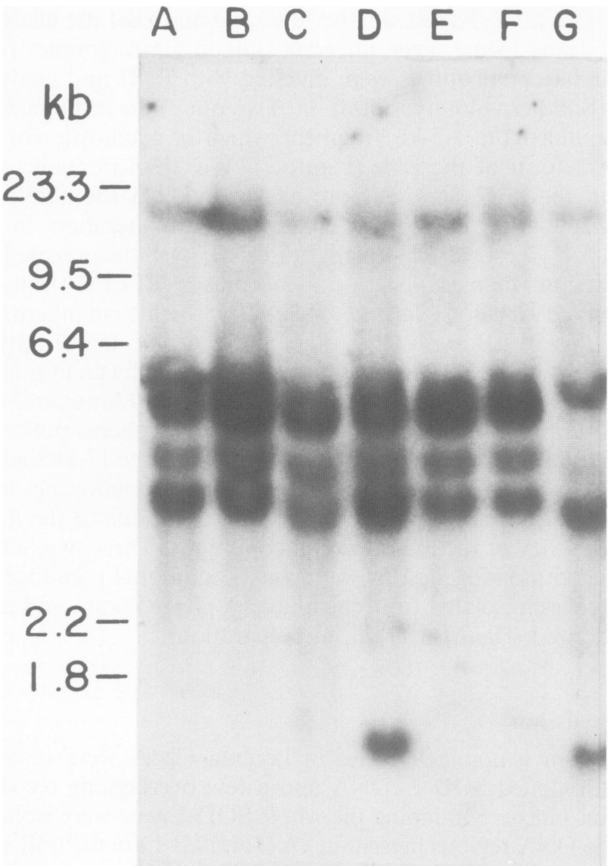

Fig. 7. Hybridization analysis revealing *Bgl*II polymorphisms in the SOD-1 gene. Each lane of the 0.8% agarose gel contained 20–30 μg of *Bgl*II digested human DNA from unrelated individuals. Transfer of DNA and hybridization with ^{32}P-labeled probe of SOD-1 cDNA described in Figure 1.

SOD-1 gene were sequenced as well as 300 bp of the 5′-flanking region and 220 bp of the 3′-flanking region. In the donor sequence of the first intron a T to C transition occurred and hence it deviates from the 5′GT. . . . AG 3′ consensus (Breathnach and Chambon, 1981; Mount, 1982). The unusual 5′-G-C donor site is not an artifact of the cloning procedure because it was detected in both λA-2 and λB-1 which are alleles of the SOD-1 gene. Since this is the only functional SOD-1 gene present in the genome there is no reason to assume that the G-C- donor site is not functional *in vivo*. In fact, we have inserted an 11-kb fragment, derived from λB-1, which contains the SOD-1 gene, into plasmid vectors carrying the bacterial phosphotranspherase gene (*neo*), and used it to transfect mouse L-cells. Many of the transformants resistant to the antibiotic G418 synthesized immunoprecipitable human SOD-1 polypeptide at relatively high efficiency indicating that indeed the G-C variant is functional (Stein and Groner, unpublished). Four violations of the 5′-G-T . . . . A-G-3′ rule have so far been reported: two within an intron, of the collagene gene (Avvedimento *et al.*, 1980) and the αA-crystallin gene (King and Piatigorsky, 1983), the other two were found in the donor sequence at the second intron of chicken (Dodgson and Engel, 1983) and duck (Erbil and Niessing, 1983) α-globin genes. In the case of chicken α-globin gene, the G-C donor site was shown to be functional *in vivo* (Fischer *et al.*, 1984). It was also reported that when a normal donor site is deleted a cryptic 5′-G-C- sequence becomes active (Wieringa *et al.*, 1983).

The 5′ transcriptional control region of the SOD-1 gene contains, in addition to the conventional 'TATA' and 'CAT' boxes, four 'GC-rich' elements which are known to be important for transcription of SV40 and a few other genes. Three of the 'GC' hexamers are direct repeats 5′-GGCGGG-3′ and the fourth is the inverted complement 5′-CCGCCC-3′. In SV40 six copies of the CCGCCC hexamer are contained within the region of the 21-bp repeats which have been shown to play an important role in both early and late transcription of the virus (Benoist and Chambon, 1981; Myers *et al.*, 1981; Lebowitz and Ghosh, 1982; Fromm and Berg 1982, 1983; Everett *et al.*, 1983; Hartzell *et al.*, 1983; Byrne *et al.*, 1983; Hansen and Sharp, 1983). Furthermore, a cellular factor which binds specifically to this region of the SV40 genome and protects three of the CCGCCC repeats was purified from HeLa cell extracts (Dynan and Tjian, 1983). This promoter-specific factor activates transcription of early and late SV40 RNAs and early BK virus RNA, but has little or no effect on a few other promoters (Dynan and Tjian, 1983, and their unpublished results). Control regions of several other genes like the herpes thymidine kinase, the mouse hypoxanthine phosphoribosyltransferase, the human and *Chlamydomonas* β-tubulin and the human HMG CoA reductase gene contain the 'GC'-rich element (McKnight and Kingsbury, 1982; Melton *et al.*, 1984; Lee *et al.*, 1983; Brunke *et al.*, 1984; Reynolds *et al.*, 1984). Repeated 'GC' motifs were also found in monkey genomic fragments capable of promoting transcription in CV-1 cells (Saffer and Singer, 1984) and the *in vivo* importance of a similar sequence in the rabbit β-globin promoter was recently reported (Dierks *et al.*, 1983). Transcription of the herpes thymidine kinase gene is dependent on the presence of two 'GC' elements at the 5′ control region; the hexanucleotide 5′-CCGCCC-3′ and its inverted complement 5′-GGCGGG-3′ (McKnight *et al.*, 1984). As indicated above, similar arrangements occur in the SOD-1 gene: two out of the four hexamers 5′-GGCGG-3′ are embedded within the 13-nucleotide inverted repeats which can form a stable stem-loop structure. Since it is generally believed that control of gene expression is achieved through interaction of regulatory proteins with specific regions of the DNA, both the 'GC' elements and the stem-loop structure might be involved in SOD-1 gene regulation.

Two mRNA species of 0.7 kb and 0.9 kb which originated from multiple polyadenylation signals at the 3′ end of the SOD-1 gene were detected (Lieman-Hurwitz *et al.*, 1982; Sherman *et al.*, 1983 and in preparation). Five poly(A) addition signals grouped in two sets were identified at the 3′ region of the gene. The first group contains two signals but only the more downstream one is utilized as a poly(A) addition signal for the 0.7-kb mRNA. The second group contains three tandem signals and here the middle one specifies the 3′ end of the 0.9-kb mRNA. Between the two polyadenylation signals of the first group there is a region of 67 bp flanked by two perfect 9-nucleotide direct repeats. It is known that transposition of DNA segments usually generates small duplications at the insertion site (Grindley and Sherratt, 1979). Therefore, it is tempting to assume that the more downstream poly(A) addition signal which specifies the end of the 0.7-kb mRNA was created by an insertion of the 67-bp fragment. Another interesting feature of the 3′ region is the presence of five ATGs followed by open reading frames that terminate with stop codons. A similar observation was previously reported for the 3′-non-coding region of chicken vimentin gene

(Zehner and Paterson, 1982). Initiation of translation of eukaryotic mRNAs takes place at an AUG near the 5′ cap. Therefore, the open reading frames in the 3′-non-coding part of the SOD-1 gene could not possibly be translated from the conventional SOD-1 mRNA species. In this context it should be mentioned that low mol. wt. RNAs were occasionally detected on Northern blots probed with ^{32}P-labeled SOD1 cDNA (Lieman-Hurwitz *et al.*, 1982; Sherman *et al.*, 1983).

The recombinant phages λA-2 and λB-1 contain the two alleles of the SOD-1 gene. They were defined by their different *Bgl*II fragments. Taking all the data in Figures 1, 2 (the genomic library) and Figure 7 together, three out of the 10 DNA samples analyzed contain the infrequent allele. Variations in DNA sequences that occur in only one of the homologous chromosomes and thus result in alteration of the length of restriction fragments (RFLPs) have been detected at various human gene loci. These sequence variants are quite prevalent, for example, in human β-globin gene locus and human albumin locus they occur once in every 100 – 200 bp and once every 85 bp, respectively (Jeffreys, 1979; Murray *et al.*, 1984). The Mendelian inheritance of the RFLPs make them important genetic markers for studies of inherited diseases (Kan and Dozy, 1978; Botstein *et al.*, 1980; Davies *et al.*, 1983; Camerino *et al.*, 1984; Drayna *et al.*, 1984).

Materials and methods

Isolation of λ clones

To isolate chromosomal DNA containing SOD-1 sequences from human bacteriophage λ libraries (Lawn *et al.*, 1978; Mory *et al.*, 1981) 10^6 phages were screened using duplicate filters and nick-translated pS61-10 cDNA clone as a probe (Lieman-Hurwitz *et al.*, 1982). The phages containing SOD-1 sequences were identified and plaque-purified.

DNA-blot hybridization

10 – 20 μg of DNA were digested with the appropriate restriction endonuclease and fractionated on a 0.8% agarose gel in TAE buffer (40 mM Tris, 20 mM Na acetate, 2 mM EDTA), containing 1 μg/ml ethidium bromide. The DNA was denatured and transferred to nitrocellulose as described by Southern (1975). Hybridizations were carried out at 42°C in 50% formamide, 5 x SSC, 5 x Denhardt, 50 mM $NaPO_4$ (pH 6.5) and 100 μg/ml of salmon sperm DNA. The filters were washed at 50°C with 1 x SSC 0.1% SDS as in Figure 7 or with 0.1 x SSC 0.1% SDS at 50°C in Figure 1.

DNA sequence analysis

DNA nucleotide sequences were determined by a combination of the chemical method of Maxam and Gilbert (1980) and dideoxynucleotide analysis (Sanger *et al.*, 1980). For the chemical degradation, restriction digest fragments were labeled either at 3′ or 5′ termini with the appropriate [^{32}P]dNTP. For dideoxynucleotide analysis, restriction fragments were subcloned into appropriately cleaved M13 DNA (Vieira and Messing, 1983).

Heteroduplex analyses

Cloned genomic and cDNA samples were mixed in a ratio of 1:2, respectively, denatured for 10 min in 0.1 M NaOH/12 mM EDTA, and neutralized by adjusting the solution to 160 mM Tris-HCl, pH 8.5/400 mM $NaClO_4$. Formamide (recrystallized three times) was added to 50% (v/v) and renaturation was allowed to proceed at 37°C for 90 min. The solution was spread on a 10% formamide hypophase and samples were prepared for electron microscopy as described by Davis *et al.*, (1971). DNA contour lengths were measured with φX174 and pUC-13 serving as internal length standards for single and double strands, respectively.

Acknowledgements

We thank Tom Maniatis for providing the human library and Jonathan Kuhn for the plasmid pHG165. This work was supported by Biotechnology General Corp., Israel, and by a Basic Research Grant No. 1-906 from the March of Dimes Birth Defects Foundation.

References

Avvedimento,V.E., Vogeli,G., Yamada,Y., Maizel,J.V.Jr., Pastan,I. and de Crombrugghe,B. (1980) *Cell*, **21**, 689-696.
Barra,D., Martini,F., Bannister,J.V., Schinina,M.E., Rotilio,G., Bannister, W.H. and Bossa,F. (1980) *FEBS Lett.*, **120**, 53-56.
Benoist,C. and Chambon,P. (1981) *Nature*, **290**, 304-310.
Botstein,D., White,R.L., Skolnick,M. and Davis,R.W. (1980) *Am. J. Hum. Genet.*, **32**, 314-331.
Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.*, **50**, 349-383.
Briggs,R.G. and Fee,J.A. (1978) *Biochim. Biophys. Acta*, **537**, 86-99.
Brunke,K.J., Anthony,J.G., Sternberg,E.J. and Weeks,D.P. (1984) *Mol. Cell. Biol.*, **4**, 1115-1124.
Burger,P.C. and Vogel,F.S. (1973) *Am. J. Pathol.*, **73**, 457-476.
Burgio,G.R., Fraccaro,M., Tiepolo,L. and Wolf,U., eds. (1981) *Trisomy 21*, published by Springer, Berlin.
Byrne,B.J., Davis,S.M., Yamaguchi,J., Bergsma,D.J. and Subramanian, K.N. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 721-725.
Camerino,G., Grzeschik,K.H., Jaye,M., De La Salle,H., Tolstosher,P., Lecocq,J.P., Heilig,R. and Mandel,J-L. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 498-502.
Cervenka,J., Gorlin,R.J. and Djavadi,G.R. (1977) *Clin. Genet.*, **11**, 119-121.
Crosti,N., Serra,A., Rigo,A. and Viglino,P. (1976) *Hum. Genet.*, **31**, 197-203.
Davis,R.W., Simon,M. and Davidson,N. (1971) *Methods Enzymol.*, **21**, 413-428.
Davies,K.E., Pearson,P.L., Harper,P.S., Murray,J.M., O'Brien,T., Sarfarazi,M. and Williamson,R. (1983) *Nucleic Acids Res.*, **11**, 2303-2312.
Dierks,P., van Ooyen,A., Cochran,M.D., Dobkin,C., Reiser,J. and Weissmann,C. (1983) *Cell*, **32**, 695-706.
Dodgson,J.B. and Engel,J.D. (1983) *J. Biol. Chem.*, **258**, 4623-4629.
Drayna,D., Davies,K., Hartley,D., Mandel,J-L., Camerinon,G., Williamson,R. and White,R. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 2836-2839.
Dynan,W.S. and Tjian,R. (1983) *Cell*, **35**, 79-87.
Ellis,W.G., McCullogh,J.R. and Corley,C.L. (1974) *Neurology*, **24**, 101-106.
Erbil,C. and Niessing,J. (1983) *EMBO J.*, **2**, 1339-1343.
Eriksson,A.W., Frants,R.R. and Jongbloet,P.H. (1975) *Am. J. Hum. Genet. (Abstr.)*, **27**, 33A.
Everett,R.D., Baty,D. and Chambon,P. (1983) *Nucleic Acid Res.*, **11**, 2447-2464.
Feaster,W.W., Kwok,L.W. and Epstein,C.J. (1977) *Am. J. Hum. Genet.*, **29**, 563-570.
Fischer,H.D., Dodgson,J.B., Hughes,S. and Engel,J.D. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 2733-2737.
Fridovich,I. (1978) *Science (Wash.)*, **201**, 875-880.
Fridovich,I. (1979) in Eichhorn,G.L. and Marzilli,L.G. (eds.), *Advances in Inorganic Biochemistry*, Elsevier/North Holland,NY, pp. 67-90.
Fromm,M. and Berg,P. (1982) *J. Mol. Appl. Genet.*, **1**, 457-481.
Fromm,M. and Berg,P. (1983) *J. Mol. Appl. Genet.*, **2**, 127-135.
Grindley,N.D.F. and Sherratt,D.J. (1979) *Cold Spring Harbor Symp. Quant. Biol.*, **43**, 1257-1261.
Hadfield,C. (1983) *Focus – RBL Technical Bulletin*, **5**, 1-5.
Hagemeijer,A. and Smith,E.M.E. (1977) *Hum. Genet.*, **38**, 15-23.
Hansen,U. and Sharp,P.A. (1983) *EMBO J.*, **2**, 2293-2303.
Hartzell,S.W., Yamaguchi,J. and Subramanian,K. (1983) *Nucleic Acids Res.*, **11**, 1601-1616.
Heston,L.L. (1977), *Science (Wash.)*, **196**, 322-323.
Jabusch,J.R., Farb,D.L., Kerschensteiner,D.A. and Deutsch,H. (1980) *Biochemistry (Wash.)*, **19**, 2310-2316.
Jeffreys,A.J. (1979) *Cell*, **18**, 1-10.
Kan,Y.W. and Dozy,A.M. (1978) *Proc. Natl. Acad. Sci. USA*, **75**, 5631-5635.
King,C.R. and Piatigorsky,J. (1983) *Cell*, **32**, 707-712.
Kozak,C.A., Lawrence,J.B. and Ruddle,F.H. (1977) *Exp. Cell Res.*, **105**, 109-117.
Krumlauf,R., Jeanpierre,M. and Young,B.D. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 2971-2975.
Lawn,R.M., Fritsch,E.F., Parker,R.C., Blake,G. and Maniatis,T. (1978) *Cell*, **15**, 1157-1174.
Lee,M.G-S., Lewis,S.A., Wilde,C.D. and Cowan,N.J. (1983) *Cell*, **33**, 477-487.
Lebowitz,P. and Ghosh,P.K. (1982) *J. Virol.*, **41**, 449-461.
Lejeune,J.M., Gautier,M. and Turpin,R. (1959) *C. R. Hebd. Seances Acad. Sci., Paris*, **248**, 1721-1722.
Lieman-Hurwitz,J., Dafni,N., Lavie,V. and Groner,Y. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 2808-2811.
Martin,G.M. (1978) *Birth Defects Orig. Article Series*, **XIV (1)**, 5-39.
Maxam,A.M. and Gilbert,W. (1980) *Methods Enzymol.*, **65**, 499-560.
McKnight,S.L. and Kingsbury,C.R. (1982) *Science (Wash.)*, **217**, 316-324.
McKnight,S.L., Kingsbury,C.R., Spence,A. and Smith,M. (1984) *Cell*, **37**, 253-262.

Melton,D.W., Konecki,D.S., Brennand,J. and Caskey,T.C. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 2147-2151.
Mory,Y., Chernajovsky,Y., Feinstein,S.I., Chen,L., Nir,U., Weissenbach,J., Malpiece,Y., Tiollais,P., Marks,D., Ladner,M., Colby,C. and Revel,M. (1981) *Eur. J. Biochem.*, **120**, 197-202.
Mount,S.M. (1982) *Nucleic Acid Res.*, **10**, 459-472.
Murray,J.C., Mills,K.A., Demopulos,C.M., Hornung,S. and Motulsky, A.G. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 3486-3490.
Myers,R.M., Rio,D.C., Robbins,A.K. and Tjian,R. (1981) *Cell*, **25**, 373-384.
Niebuhr,E. (1974) *Hum. Genet.*, **21**, 99-100.
Owens,D., Dawson,J.C. and Losin,S. (1971) *Am. J. Ment. Defic.*, **75**, 606-612.
Pfeiffer,R.A., Kessel,E.K. and Soer,K.H. (1977) *Clin. Genet.*, **11**, 207-213.
Philip,T., Fraisse,J., Sinet,P.M., Lauras,B., Roberts,J.M. and Freycon,F. (1978) *Cytogenet. Cell Genet.*, **22**, 521-523.
Poissonnier,M., St. Paul,B., Dutrillaux,B., Chassegne,M., Gruyer,M. and Bligniers-Strauk,G. (1976) *Ann. Genet.*, **19**, 69-73.
Reynolds,G.A., Basu,S.K., Osborne,T.F., Chin,D.J., Gil,G., Brown,M.S., Goldstein,J.L. and Luskey,K.L. (1984) *Cell*, **38**, 275-285.
Saffer,J.D. and Singer,M.F. (1984) *Nucleic Acids Res.*, **12**, 4769-4788.
Sanger,F., Coulson,A.R., Barrel,B.G., Smith,A.J.H. and Roe,B.A. (1980) *J. Mol. Biol.*, **143**, 161-178.
Sherman,L., Dafni,N., Lieman-Hurwitz,J. and Groner,J. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 5465-5469.
Sinet,P.M., Allard,D., Lejeune,J. and Jerome,H. (1974) *C. R. Hebd. Seances Acad. Sci., Paris*, **278**, 3267-3270.
Sinet,P.M., Coutourier,J., Dutrillaux,B., Poissonnier,M., Raoul,O., Rethore,M.O., Allard,D., Lejeune,J. and Jerome,H. (1976) *Exp. Cell Res.*, **97**, 47-55.
Sinet,P.M. (1982) *Ann. N.Y. Acad. Sci.*, **396**, 83-94.
Solitaire,G. and Lamarche,J. (1966) *Am. J. Ment. Dis.*, **70**, 840-848.
Southern,E.M. (1975) *J. Mol. Biol.*, **98**, 503-517.
Summit,R.L. (1981) in de la Cruz,F.F. and Gerald,P.S. (eds.), *Trisomy 21 (Down Syndrome) Research Perspective*, University Park Press, Baltimore, pp. 225-235.
Tan,Y.H., Tischfield,J. and Ruddle,F.H. (1973) *J. Exp. Med.*, **137**, 317-330.
Tinoco,I.Jr., Borer,P.N., Dengler,B., Levine,M.D., Uhlenbeck,O.C., Crothers,D.M. and Gralla,J. (1973) *Nature New Biol.*, **246**, 40-41.
Tsukino,R., Tsuda,N., Dezawa,T., Ishü,T. and Koike,M. (1980) *J. Mol. Genet.*, **17**, 144-155.
Vieira,J. and Messing,J. (1983) *Gene*, **19**, 259-268.
Wieringa,B., Meyer,F., Reiser,J. and Weissmann,C. (1983) *Nature*, **301**, 38-43.
Williams,J.D., Summut,R.L., Martens,P.R. and Kimbrell,R.A. (1975) *Am. J. Hum. Genet.*, **27**, 478-481.
Zehner,Z.E. and Paterson,B. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 911-915.

Received on 19 September 1984; revised on 29 October 1984