MOLECULAR AND CELLULAR BIOLOGY, Apr. 1992, p. 1687–1697
0270-7306/92/041687-11$02.00/0

Vol. 12, No. 4

The Translocation (6;9), Associated with a Specific Subtype of Acute Myeloid Leukemia, Results in the Fusion of Two Genes, *dek* and *can*, and the Expression of a Chimeric, Leukemia-Specific *dek-can* mRNA

MARIEKE VON LINDERN, MAARTEN FORNEROD, SJOZÈF VAN BAAL, MARTINE JAEGLE, TON DE WIT, ARJAN BUIJS, AND GERARD GROSVELD*

Department of Cell Biology and Genetics, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Received 11 December 1991/Accepted 24 January 1992

The translocation (6;9) is associated with a specific subtype of acute myeloid leukemia (AML). Previously, it was found that breakpoints on chromosome 9 are clustered in one of the introns of a large gene named *Cain* (*can*). cDNA probes derived from the 3′ part of *can* detect an aberrant, leukemia-specific 5.5-kb transcript in bone marrow cells from t(6;9) AML patients. cDNA cloning of this mRNA revealed that it is a fusion of sequences encoded on chromosome 6 and 3′ *can*. A novel gene on chromosome 6 which was named *dek* was isolated. In *dek* the t(6;9) breakpoints also occur in one intron. As a result the *dek-can* fusion gene, present in t(6;9) AML, encodes an invariable *dek-can* transcript. Sequence analysis of the *dek-can* cDNA showed that *dek* and *can* are merged without disruption of the original open reading frames and therefore the fusion mRNA encodes a chimeric DEK-CAN protein of 165 kDa. The predicted DEK and CAN proteins have molecular masses of 43 and 220 kDa, respectively. Sequence comparison with the EMBL data base failed to show consistent homology with any known protein sequences.

Defined karyotypic aberrations are associated with specific subtypes of leukemia. Detailed molecular characterization of these aberrations may identify genes involved in leukemogenesis and in the precise regulation of proliferation and differentiation in the hematopoietic system. Translocations are the best-studied chromosomal abnormalities. As the result of a translocation, the function or activity of oncogenes located at or near the translocation breakpoint is altered. In myeloid leukemia three translocation breakpoints have been cloned and analyzed at the molecular level.

The two best studied, t(9;22) in chronic myeloid leukemia (27, 43) and t(15;17) in acute promyelocytic leukemia (2, 8, 12), result in the formation of chimeric genes that encode fusion proteins. In chronic myeloid leukemia this is a BCR-ABL protein that has an enhanced tyrosine kinase activity (34, 49) directly responsible for its in vivo tumorigenic potential (14, 25). In acute promyelocytic leukemia a PML-RARα fusion protein that represents an altered transcription factor (16, 33) is found.

The third translocation is the t(6;9) (p23;q34), found in a specific subtype of acute myeloid leukemia (AML) (1, 39, 41). This leukemia is characterized by a poor prognosis, affects young adults, and is classified mostly as M2 or M4 and rarely as M1 (according to the French-American-British classification of AML). A region on chromosome 9 situated 360 kb downstream of the c-*abl* gene was cloned and analyzed. It was found that breakpoints were clustered in a region of 8 kb in five patients, four with t(6;9) AML and one with acute undifferentiated leukemia (AUL) (47). Through cDNA cloning this region could be identified as one of the introns of a large gene (>100 kb) encoding a 7-kb transcript. This intron was named *icb*-9; the intron containing the breakpoints on chromosome 9 and situated in the middle of a gene named *Cain* (*can*). The 3′ part of *can* is translocated to the 6p− chromosome, and only 3′ *can* probes detect an additional, leukemia-specific 5.5-kb transcript in bone marrow cells from t(6;9) AML patients. No additional transcripts were detected with 5′ *can* probes. The breakpoint region on chromosome 6p23 was isolated from a genomic λEMBL3 library constructed of bone marrow DNA from one of the t(6;9) patients. An area of 40 kb of chromosome 6 DNA was cloned in overlapping phages. Southern blot analysis showed that chromosomal breakpoints t(6;9) AML patients are clustered in a relatively small region of 12 kb.

This article reports the cloning of a cDNA representing the 5.5-kb aberrant transcript specific for t(6;9) AML; the isolation of a novel gene, *dek*, on chromosome 6p23; and the sequence analysis of both *can* and *dek* cDNAs.

MATERIALS AND METHODS

Northern (RNA) blotting. Patient material and cell lines used were described previously (47, 48). RNA of mouse tissue was isolated from BCBA mice. RNA was isolated by either the guanidinium isothiocyanate (11) or the LiCl-Ureum method (5). Total RNA was electrophoresed and blotted as described by Fourney et al. (20). Equal amounts of rRNA were loaded; before the samples were loaded on a denaturing gel, 5% of each sample was loaded on a nondenaturing agarose gel to estimate the amount of rRNA and to adapt the sample quantity if necessary. Northern blots were hybridized in 10% dextran (40). Northern blots of mouse tissues were hybridized with human probes with 3× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate) at 65°C, and filters were washed with 1× SSC at 65°C for *dek* probes and with 0.3× SSC at 65°C for *can* probes. Probes were labelled by the method of Feinberg and Vogelstein (19).

cDNA cloning. One hundred micrograms of total RNA from patient DK was heat denatured and annealed to 10 μg

* Corresponding author.

of a 21-mer, 5'GAAGGACTAGGTGCACCATGT3', at 55°C. First-strand synthesis was done with avian reverse transcriptase (26). Second-strand synthesis was done according to the RNaseH method (24). The DNA was blunt ended with T4 polymerase and treated with *Eco*RI methylase (Sigma). *Eco*RI linkers were ligated onto the cDNA with T4 ligase and RNA ligase (40), and after *Eco*RI digestion, the cDNA was size selected on a Sephacryl S-1000 column. cDNA larger than 1 kb was ligated into the *Eco*RI site of λgt10 (31). Phage DNA was packaged by using packaging extracts (GIGA gold; Stratagene). PFU (19×10^6) were generated, of which only 10% contained inserts, estimated by analysis of randomly picked phages. The other 90% most likely contained linker sequences. The human testis cDNA library in λgt11 was purchased from Clonetech (Palo Alto, Calif.). The CMLO λEMBL3 library was described by Hermans et al. (28).

Sequence determination and analysis. Restriction fragments of cDNA clones were subcloned in M13. Overlapping cDNA sequences on both strands were determined by dideoxy sequencing (42). Initially, M13 primers were used; when no suitable restriction sites were present a primer was generated on the basis of already available cDNA sequence. To establish intron-exon borders, genomic fragments containing the exon of interest were subcloned into M13 and a primer near the putative intron-exon border was generated to prime the sequence reaction. Sequences were analyzed with the computer program Microgenie, and the EMBL data base was used to search for homologous sequences at both the nucleotide and amino acid levels.

Cloning of the 3' end of *can*. Thirty micrograms of total RNA of bone marrow cells from AUL patient SE was heat denatured, and first-strand cDNA was synthesized with avian reverse transcriptase by using 100 pmol of the 35-mer 5'GTCGCGAATTCGTCGACGCGTTTTTTTTTTTTTTT3' as a primer (21, 28). Excess primer was removed by isopropanol precipitation. One hundredth of the cDNA reaction was amplified by using *Taq* polymerase (Perkin-Elmer Cetus) and the primers 5'GTCGCGAATTCGTCGACGCG3' and 5'GCCTTTGGATCCCTGGGACCAACCGC3'. The latter primer is located 180 bp upstream of the poly(A) signal in *can* cDNA. The amplified fragment of 230 bp was sequenced by using a protocol for direct sequencing of fragments produced by an asymmetric polymerase chain reaction (32).

Nucleotide sequence accession numbers. The nucleotide sequence data reported in this article have been submitted to the EMBL, Genbank, and DDBJ nucleotide sequence data bases under accession numbers X64228 (*can*) and X64229 (*dek*).

RESULTS

Analysis of the *can* gene and transcript. As reported previously, a nearly full-length *can* cDNA was isolated in the overlapping cDNA clones hXT23, hXT37, hXT54, and hXT65 (47). Originally, cDNA clone hX8 was thought to represent the 5' part of the *can* mRNA. However, a more detailed mapping analysis showed that the 5' part of hX8 does not belong to the *can* gene and is in fact not even located on chromosome 9 (data not shown). Therefore, hX8 must be considered a cloning artifact. As several (11) independent cDNA clones appeared to have 5' ends mapping close to the 5' end of hXT23, we assumed that the 5' end of the latter clone maps in the vicinity of the *can* mRNA cap site.

The genomic map of *can*, reported previously, extended over 70 kb but did not include the 3' part of the *can* gene. Therefore, cDNA clone hXT65 was used to screen a genomic λEMBL3 library, and many hybridizing phages were isolated. Clones Al1F10.6, Al1F10.2, Al1F10.8, and Al1F10.12 were selected since they covered the largest stretch of DNA, and they were analyzed in more detail. As indicated in Fig. 1A, a gap is still present between Al1F10.8 and Al1F10.12. The total amount of *can* sequences cloned in phages is 130 kb. Since the gene is located on a *Bss*HII fragment of 170 kb (47) and no *Bss*HII site is present in Al1F10.12, it was deduced from Fig. 1A that the gap between Al1F10.8 and Al1F10.12 can range between 1 and 40 kb. Restriction enzyme fragments that contain exons were determined by hybridization of Southern blots containing *Eco*RI, *Bam*HI, and *Hin*dIII digests of the phages with *can* cDNA clone hXT65 (Fig. 1C).

The overlapping *can* cDNA clones were sequenced and appeared to contain a large open reading frame (ORF) of 6,270 nucleotides (nt) encoding a putative protein of 220 kDa (Fig. 2). This ORF starts in clone hXT23 and ends in clone hXT65. A 700-bp *Hin*dIII-*Pst*I fragment of phage Al1F3, in which the *Bss*HII site is located (Al1F3E4HP), was also sequenced. Figure 3A shows that the sequence of Al1F3E4HP is colinear with hXT23 up to its 5' end. Other cDNA clones have 5' ends mapping near the 5' end of hXT23. Whether this region contains *can* promoter sequences has to be tested. At the 5' end, the *can* cDNA contains ATG start codons at positions 95 to 97, 107 to 109, and 115 to 117. The sequence around the codon at position 95 is concordant with the consensus sequence postulated by Kozak (35), which suggests that this methionine is probably the start of the CAN protein. The first stop codon in this frame is at position 6365. The sequence of cDNA clone hXT65 ends immediately 3' of what appeared to be a variant polyadenylation signal: ATTAAA (nt 6562 to 6567). As no poly(A) tail was present in this clone, the 3' end of the *can* transcript was amplified by using the protocol for rapid amplification of cDNA ends (21) from a position 180 bp 5' of the poly(A) signal to the poly(A) tail. The sequence of this amplified fragment showed that the poly(A) tail starts 16 nt downstream of the ATTAAA signal. The 3' end of hXT65 hybridized to genomic λEMBL3 phage Al1F10.12. Sequence analysis showed that the 3' exon of *can* is present in this phage. Its sequence is colinear with the cDNA sequence down to the poly(A) tail (Fig. 3B).

Since previous mapping data localized the t(6;9) breakpoints in the middle of cDNA clone hXT37 (Fig. 1C), the breakpoints must occur within the *can* ORF. To exactly localize the position of *icb*-9 within the ORF, genomic clones were used to sequence the intron-exon borders delineating this intron. This showed that the translocation breakpoints occur between codons 812 and 813 (nt 2530 to 2531) in the ORF of the *can* mRNA (Fig. 2 and 3C). Because of the translocation, 4,053 nt of the *can* cDNA are encoded on the 6p− chromosome. As a consequence, cDNA probes located within these 4,053 nt recognize a specific 5.5-kb transcript in bone marrow cells from t(6;9) AML patients (47).

Cloning the *dek-can* hybrid cDNA. To resolve the identity of the t(6;9) AML-specific 5.5-kb mRNA, a primed cDNA library was constructed by using total RNA of bone marrow from t(6;9) patient DK and a 21-nt primer mapping 800 bp downstream of the translocation breakpoint in the *can* cDNA (Fig. 1C). Part of the library (2×10^6 PFU) was screened with a 360-bp *Bam*HI-*Rsa*I (hXT37BR) fragment, indicated in Fig. 1C. Two clones (DK1 and DK2 [1.3 and 1.5 kb, respectively]) were isolated and characterized. They

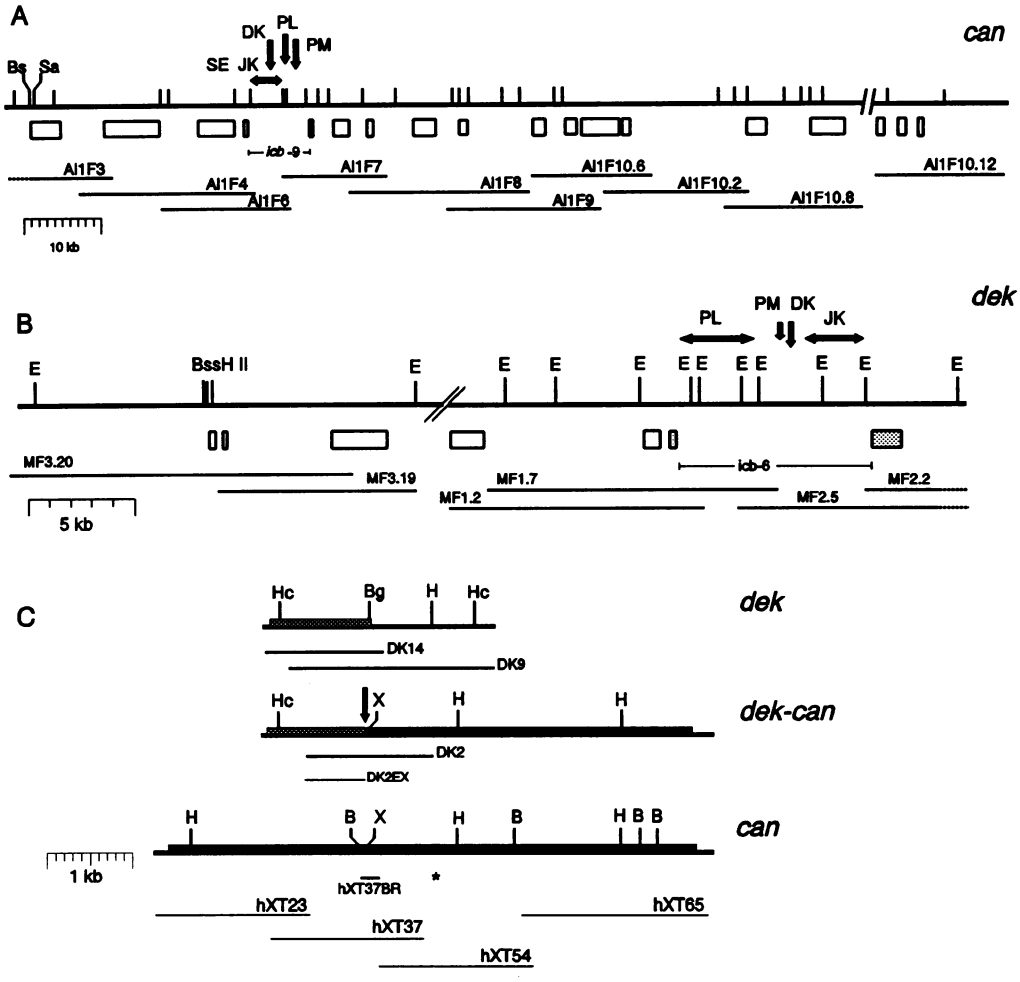

FIG. 1. Restriction maps of the *can* and *dek* genes and *dek*, *dek-can*, and *can* cDNAs. (A) Genomic map of the *can* gene. Vertical lines represent *Eco*RI sites. Open boxes represent restriction enzyme fragments hybridizing to *can* cDNA probes. The positions of the breakpoints of chromosomes from t(6;9) AML patients DK, PM, JK, and PL and AUL patient SE are indicated by arrows. They are all located in *icb*-9. Below the map, isolated genomic phages are depicted. Al1F3, Al1F4, Al1F6, Al1F7, Al1F8, and Al1F9 were reported previously. Al1F10.6, Al1F10.2, Al1F10.8, and Al1F10.12 were isolated by using cDNA clone hXT65 as probe. The gap between Al1F10.8 and Al1F10.12 is at maximum 40 kb. The scale is indicated in kilobases. (B) Genomic map of *dek*. Open boxes indicate restriction fragments hybridizing to cDNA probes; these fragments were delimited by various restriction enzyme sites not shown in this map. Stippled boxes are mapped exons. The positions of the breakpoints of chromosomes from t(6;9) AML patients DK, PM, PL, and JK are indicated by arrows; they are all located in *icb*-6. MF3.20, MF3.19, MF1.2, MF1.7, MF2.5, and MF2.2 are λEMBL3 phages from which the map has been deduced. The gap between MF3.19 and MF1.2 is estimated to be only a few kilobases. (C) Restriction maps of the cDNAs of *dek*, *dek-can*, and *can* are depicted. A scale for the cDNA maps is indicated. Arrows indicate the position of the breakpoints. The ORF of *dek* is indicated by a cross-hatched bar, and the ORF of *can* is indicated by a solid bar on top of the lines that indicate the cDNAs. The chimeric cDNA DK2 has been isolated from a primed cDNA library that was made with a primer, indicated by an asterisk. This library was screened with probe hXT37BR. DK9 and DK14 are *dek* cDNAs isolated with probe DK2EX from a λgt11 cDNA library derived from human testis RNA. Abbreviations: B, *Bam*HI; Bg, *Bgl*II; E, *Eco*RI; H, *Hin*dIII; Hc, *Hin*cII; X, *Xba*I.

appeared to be colinear with *can* cDNA from the primer at the 3′ end of the cDNA clones, exactly up to the 5′ end of the exon flanking *icb*-9 at its 3′ side. Upstream of this point, both clones are identical but deviate completely from the *can* sequence. To determine the chromosomal origin of these sequences, a 5′ DK2 fragment (probe DK2EX, a 700-bp *Eco*RI-*Xba*I fragment [Fig. 1C]) was hybridized to a Southern blot containing DNA of a hybrid cell panel with the segregated translocation chromosomes involved in the t(6;9) (48). The probe hybridized to DNA of cell lines containing chromosome 6 and 6p− (results not shown). The same probe was hybridized to a Northern blot containing RNA of HeLa cells, hematopoietic cell lines (Daudi, HL60, KG1, and K562), and bone marrow cells from t(6;9) AML patient DK and AUL patient SE. This revealed the presence of a 2.7-kb transcript in all lanes and an additional 5.5-kb transcript in the t(6;9) AML patient bone marrow sample (Fig. 4). This 5.5-kb transcript is identical in size to the aberrant transcript detected with 3′ *can* probes in a sample from this patient (47). These results proved that sequences encoded by a gene on chromosome 6 are present in the t(6;9) AML-specific 5.5-kb transcript, which is thus identified as a chimeric mRNA.

It is noteworthy that in a sample from AUL patient SE no aberrant transcript was detected by the chromosome 6 probe, while hybridization with 3′ *can* probes clearly de-

```
                                                                                                    M  G  D  E  M  D  A  M  I      9
   1 AGGGGAGGAAGTTTGCTGTCGAGCGGCCTGGGTTCCGTGGGCAAGGCGTGGGTGGCAGCGTTGGCTCGTTCGACGACACACTGAGGGCGGCGCGATGGGAGACGAGATGGATGCCATGAT

      P  E  R  E  M  K  D  F  Q  F  R  A  L  K  K  V  R  I  F  D  S  P  E  E  L  P  K  E  R  S  S  L  L  A  V  S  N  K  Y  G     49
 121 TCCCGAGCGGGAGATGAAGGATTTTCAGTTTAGAGCGCTAAAGAAGGTGAGAATCTTTGACTCCCCTGAGGAATTGCCCAAGGAACGCTCGAGTCTGCTTGCTGTGTCCAACAAATATGG

      L  V  F  A  G  G  A  S  G  L  Q  I  F  P  T  K  N  L  L  I  Q  N  K  P  G  D  D  P  N  K  I  V  D  K  V  Q  G  L  L  V     89
 241 TCTGGTCTTCGCTGGTGGAGCCAGTGGCTTGCAGATTTTTCCTACTAAAAATCTTCTTATTCAAAATAAACCCGGAGATGATCCCAACAAAATAGTTGATAAAGTCCAAGGCTTGCTAGT

      P  M  K  F  P  I  H  H  L  A  L  S  C  D  N  L  T  L  S  A  C  M  M  S  S  E  Y  G  S  I  I  A  F  F  D  V  R  T  F  S    129
 361 TCCTATGAAATTCCCAATCCATCACCTGGCCTTGAGCTGTGATAACCTCACACTCTCTGCGTGCATGATGTCCAGTGAATATGGTTCCATTATTGCTTTTTTTGATGTTCGCACATTCTC

      N  E  A  K  Q  Q  K  R  P  F  A  Y  H  K  L  L  K  D  A  A  G  M  V  I  D  M  K  W  N  P  T  V  P  S  M  V  A  V  C  L    169
 481 AAATGAGGCTAAACAGCAAAAACGCCCATTTGCCTATCATAAGCTTTTGAAAGATGCAGCAGGCATGGTGATTGATATGAAGTGGAACCCCACTGTCCCCTCCATGGTGGCAGTTTGTCT

      A  D  G  S  I  D  V  L  Q  V  T  E  T  V  K  V  C  A  T  L  P  S  T  V  A  V  T  S  V  C  W  S  P  K  G  K  Q  L  A  V    209
 601 GGCTGATGGTAGTATTGATGTCCTGCAAGTCACGGAAACAGTGAAAGTATGTGCAACTCTTCCTTCCACGGTAGCAGTAACCTCTGTGTGCTGGAGCCCCAAAGGAAAGCAGCTGGCAGT

      G  K  Q  N  G  T  V  V  Q  Y  L  P  T  L  Q  E  K  K  V  I  P  C  P  P  F  Y  E  S  D  H  P  V  R  V  L  D  V  L  W  I    249
 721 GGGAAAACAGAATGGAACTGTGGTCCAGTATCTTCCTACTTTGCAGGAAAAAAAAGTCATTCCTTGTCCTCCGTTTTATGAGTCAGATCATCCTGTCAGAGTTCTGGATGTGCTGTGGAT

      G  T  Y  V  F  A  I  V  Y  A  A  A  D  G  T  L  E  T  S  P  D  V  V  M  A  L  L  P  K  K  E  E  K  H  P  E  I  F  V  N    289
 841 TGGTACCTACGTCTTCGCCATAGTGTATGCTGCTGCAGATGGGACCCTGGAAACGTCTCCAGATGTGGTGATGGCTCTACTACCGAAAAAAGAAGAAAAGCACCCAGAGATATTTGTGAA

      F  M  E  P  C  Y  G  S  C  T  E  R  Q  H  H  Y  Y  L  S  Y  I  E  E  W  D  L  V  L  A  A  S  A  A  S  T  E  V  S  I  L    329
 961 CTTTATGGAGCCCTGTTATGGCAGCTGCACGGAGAGACAGCATCATTACTACCTCAGTTACATTGAGGAATGGGATTTAGTGCTGGCAGCATCTGCGGCTTCAACAGAAGTTAGTATCCT

      A  R  Q  S  D  Q  I  N  W  E  S  W  L  L  E  D  S  S  R  A  E  L  P  V  T  D  K  S  D  D  S  L  P  M  G  V  V  V  D  Y    369
1081 TGCTCGACAAAGTGATCAGATTAATTGGGAATCTTGGCTACTGGAGGATTCTAGTCGAGCTGAATTGCCTGTGACAGACAAGAGTGATGACTCCTTGCCCATGGGAGTTGTCGTAGACTA

      T  N  Q  V  E  I  T  I  S  D  E  K  T  L  P  P  A  P  V  L  M  L  L  S  T  D  G  V  L  C  P  F  Y  M  I  N  Q  N  P  G    409
1201 TACAAACCAAGTGGAAATCACCATCAGTGATGAAAAGACTCTTCCTCCTGCTCCAGTTCTCATGTTACTTTCAACAGATGGTGTGCTTTGTCCATTTTATATGATTAATCAAAATCCTGG

      V  K  S  L  I  K  T  P  E  R  L  S  L  E  G  E  R  Q  P  K  S  P  G  S  T  P  T  T  P  T  S  S  Q  A  P  Q  K  L  D  A    449
1321 GGTTAAGTCTCTCATCAAAACACCAGAGCGACTTTCATTAGAAGGAGAGCGACAGCCCAAGTCACCAGGAAGTACTCCCACTACCCCAACCTCCTCTCAAGCCCCACAGAAACTGGATGC

      S  A  A  A  A  P  A  S  L  P  P  S  S  P  A  A  P  I  A  T  F  S  L  L  P  A  G  G  A  P  T  V  F  S  F  G  S  S  S  L    489
1441 TTCTGCAGCTGCAGCCCCTGCCTCTCTGCCACCTTCATCACCTGCTGCTCCCATTGCCACTTTTTCTTTGCTTCCTGCTGGTGGAGCCCCCACTGTGTTCTCCTTTGGTTCTTCATCTTT

      K  S  S  A  T  V  T  G  E  P  P  S  Y  S  S  G  S  D  S  S  K  A  A  P  G  P  G  P  S  T  F  S  F  V  P  P  S  K  A  S    529
1561 GAAGTCATCTGCTACGGTCACTGGGGAGCCCCCTTCATATTCCAGTGGCTCCGACAGCTCCAAAGCAGCCCCAGGCCCTGGCCCATCAACCTTCTCTTTTGTTCCCCCTTCTAAAGCCTC

      L  A  P  T  P  A  A  S  P  V  A  P  S  A  A  S  F  S  F  G  S  S  G  F  K  P  T  L  E  S  T  P  V  P  S  V  S  A  P  N    569
1681 CCTAGCCCCCACCCCTGCAGCGTCTCCTGTGGCTCCATCAGCTGCTTCATTCTCCTTTGGATCATCTGGTTTTAAGCCTACCCTGGAAAGCACACCAGTGCCAAGTGTGTCTGCTCCAAA

      I  A  M  K  S  S  F  P  P  S  T  S  A  V  K  V  N  L  S  E  K  F  T  A  A  A  T  S  T  P  V  S  S  S  Q  S  A  P  P  M    609
1801 TATAGCAATGAAGTCCTCCTTCCCACCCTCAACCTCTGCTGTCAAAGTCAACCTTAGTGAAAAGTTTACTGCTGCAGCTACCTCTACTCCTGTTAGTAGCTCCCAGAGCGCACCCCCGAT

      S  P  F  S  S  A  S  K  P  A  A  S  G  P  L  S  H  P  T  P  L  S  A  P  P  S  S  V  P  L  K  S  S  V  L  P  S  P  S  G    649
1921 GTCGCCATTCTCTTCTGCCTCCAAGCCAGCTGCTTCTGGACCACTCAGCCACCCCACGCCTCTCTCAGCACCACCTAGTTCCGTGCCATTGAAGTCCTCAGTCTTGCCCTCACCATCAGG

      R  S  A  Q  G  S  S  S  P  V  P  S  M  V  Q  K  S  P  R  I  T  P  P  A  A  K  P  G  S  P  Q  A  K  S  L  Q  P  A  V  A    689
2041 ACGATCTGCTCAGGGCAGTTCAAGCCCAGTGCCCTCAATGGTACAGAAATCACCCAGGATAACCCCTCCAGCGGCAAAGCCAGGCTCTCCCCAGGCAAAGTCACTTCAGCCTGCTGTTGC

      E  K  Q  G  H  Q  W  K  D  S  D  P  V  M  A  G  I  G  E  E  I  A  H  F  Q  K  E  L  E  E  L  K  A  R  T  S  K  A  C  F    729
2161 AGAAAAGCAGGGACATCAGTGGAAAGATTCAGATCCTGTAATGGCTGGAATTGGGGAGGAGATTGCACACTTTCAGAAGGAGTTGGAAGAGTTAAAAGCCCGAACTTCCAAAGCCTGTTT

      Q  V  G  T  S  E  E  M  K  M  L  R  T  E  S  D  D  L  H  T  F  L  L  E  I  K  E  T  T  E  S  L  H  G  D  I  S  S  L  K    769
2281 CCAAGTGGGCACTTCTGAGGAGATGAAGATGCTGCGAACAGAATCAGATGACTTGCATACCTTTCTTTTGGAGATTAAAGAGACCACAGAGTCGCTTCATGGAGATATAAGTAGCCTGAA

      T  T  L  L  E  G  F  A  G  V  E  E  A  R  E  Q  N  E  R  N  R  D  S  G  Y  L  H  L  L  Y  K  R  P  L  D  P  K  S  E  A    809
2401 AACAACTTTACTTGAGGGCTTTGCTGGTGTTGAGGAAGCCAGAGAACAAAATGAAAGAAATCGTGACTCTGGTTATCTGCATTTGCTTTATAAAAGACCACTGGATCCCAAGAGTGAAGC

      Q  L  Q  E  I  R  R  L  H  Q  Y  V  K  F  A  V  Q  D  V  N  D  V  L  D  L  E  W  D  Q  H  L  E  Q  K  K  K  Q  R  H  L    849
2521 TCAGCTTCAGGAAATTCGGCGCCTTCATCAGTATGTGAAATTTGCTGTCCAAGATGTGAATGATGTTCTAGACTTGGAGTGGGATCAGCATCTGGAACAAAAGAAAAAACAAAGGCACCT
               ▲
      L  V  P  E  R  E  T  L  F  N  T  L  A  N  N  R  E  I  I  N  Q  Q  R  K  R  L  N  H  L  V  D  S  L  Q  Q  L  R  L  Y  K    889
2641 GCTTGTGCCAGAGCGAGAGACACTGTTTAACACCCTAGCCAACAATCGGGAAATCATCAACCAACAGAGGAAGAGGCTGAATCACCTGGTGGATAGTCTTCAGCAGCTCCGCCTTTACAA

      Q  T  S  L  W  S  L  S  S  A  V  P  S  Q  S  S  I  H  S  F  D  S  D  L  E  S  L  C  N  A  L  L  K  T  T  I  E  S  H  T    929
2761 ACAGACTTCCCTGTGGAGCCTGTCCTCGGCTGTTCCTTCCCAGAGCAGCATTCACAGTTTTGACAGTGACCTGGAAAGCCTGTGCAATGCTTTGTTGAAAACCACCATAGAATCTCACAC

      K  S  L  P  K  V  P  A  K  L  S  P  M  K  Q  A  Q  L  R  N  F  L  A  K  R  K  T  P  P  V  R  S  T  A  P  A  S  L  S  R    969
2881 CAAATCCTTGCCCAAAGTACCAGCCAAACTGTCCCCCATGAAACAGGCACAACTGAGAAACTTCTTGGCCAAGAGGAAGACCCCACCAGTGAGATCCACTGCTCCAGCCAGCCTGTCTCG

      S  A  F  L  S  Q  R  Y  Y  E  D  L  D  E  V  S  S  T  S  S  V  S  Q  S  L  E  S  E  D  A  R  T  S  C  K  D  D  E  A  V   1009
3001 ATCAGCCTTTCTGTCTCAGAGATATTATGAAGACTTGGATGAAGTCAGCTCAACGTCATCTGTCTCCCAGTCTCTGGAGAGTGAAGATGCACGGACGTCCTGTAAAGATGACGAGGCAGT

      V  Q  A  P  R  H  A  P  V  V  R  T  P  S  I  Q  P  S  L  L  P  H  A  A  P  F  A  K  S  H  L  V  H  G  S  S  P  G  V  M   1049
3121 GGTTCAGGCCCCTCGGCACGCCCCCGTGGTTCGCACTCCTTCCATCCAGCCCAGTCTCTTGCCCCATGCAGCACCTTTTGCTAAATCTCACCTGGTTCATGGTTCTTCACCTGGTGTGAT

      G  T  S  V  A  T  S  A  S  K  I  I  P  Q  G  A  D  S  T  M  L  A  T  K  T  V  K  H  G  A  P  S  P  S  H  P  I  S  A  P   1089
3241 GGGAACTTCAGTGGCTACATCTGCTAGCAAAATTATTCCTCAAGGGGCCGATAGCACAATGCTTGCCACGAAAACCGTGAAACATGGTGCACCTAGTCCTTCCCACCCCATCTCAGCCCC

      Q  Q  L  A  A  A  A  L  R  R  Q  M  A  S  Q  A  P  A  V  N  T  L  T  E  S  T  L  K  N  V  P  Q  V  V  N  V  Q  E  L  K   1129
3361 GCAGCAGCTGGCCGCAGCAGCACTCAGGCGGCAGATGGCCAGTCAGGCACCAGCTGTAAACACTTTGACTGAATCAACGTTGAAGAATGTCCCTCAAGTGGTAAATGTGCAGGAATTGAA

      N  N  P  A  T  P  S  T  A  M  G  S  S  V  P  Y  S  T  A  K  T  P  H  P  V  L  T  P  V  A  A  N  Q  A  K  Q  G  S  L  I   1169
3481 GAATAACCCTGCAACCCCTTCTACAGCCATGGGTTCTTCAGTGCCCTACTCCACAGCCAAAACACCTCACCCAGTGTTGACCCCAGTGGCTGCTAACCAAGCCAAGCAGGGGTCTCTAAT

      N  S  L  K  P  S  G  P  T  P  A  S  G  Q  L  S  S  G  D  K  A  S  G  T  A  K  I  E  T  A  V  T  S  T  P  S  A  S  G  Q   1209
3601 AAATTCCCTTAAGCCATCTGGGCCTACACCAGCATCCGGTCAGTTATCATCTGGTGACAAAGCTTCAGGGACAGCCAAGATAGAAACAGCTGTGACTTCAACCCCATCTGCTTCTGGGCA

      F  S  K  P  F  S  F  S  P  S  G  T  G  F  N  F  G  I  I  T  P  T  P  S  S  N  F  T  A  A  Q  G  A  T  P  S  T  K  E  S   1249
3721 GTTCAGCAAGCCTTTCTCATTTTCTCCATCAGGGACTGGCTTTAATTTTGGGATAATCACACCAACACCGTCTTCTAATTTCACTGCTGCACAAGGGGCAACACCCTCCACTAAAGAGTC

      S  Q  P  D  A  F  S  S  G  G  G  S  K  P  S  Y  E  A  I  P  E  S  S  P  P  S  G  I  T  S  A  S  N  T  T  P  G  E  P  A   1289
3841 AAGCCAGCCGGACGCATTCTCATCTGGTGGGGGAAGCAAACCTTCTTATGAGGCCATTCCTGAAAGCTCACCTCCCTCAGGAATCACATCCGCATCAAACACCACCCCAGGAGAACCTGC

      A  S  S  S  R  P  V  A  P  S  G  T  A  L  S  T  T  S  S  K  L  E  T  P  P  S  K  L  G  E  L  L  F  P  S  S  L  A  G  E   1329
3961 CGCATCTAGCAGCAGACCTGTGGCACCTTCTGGAACTGCTCTTTCCACCACCTCTAGTAAGCTGGAAACCCCACCGTCCAAGCTGGGAGAGCTTCTGTTTCCAAGTTCTTTGGCTGGAGA
```

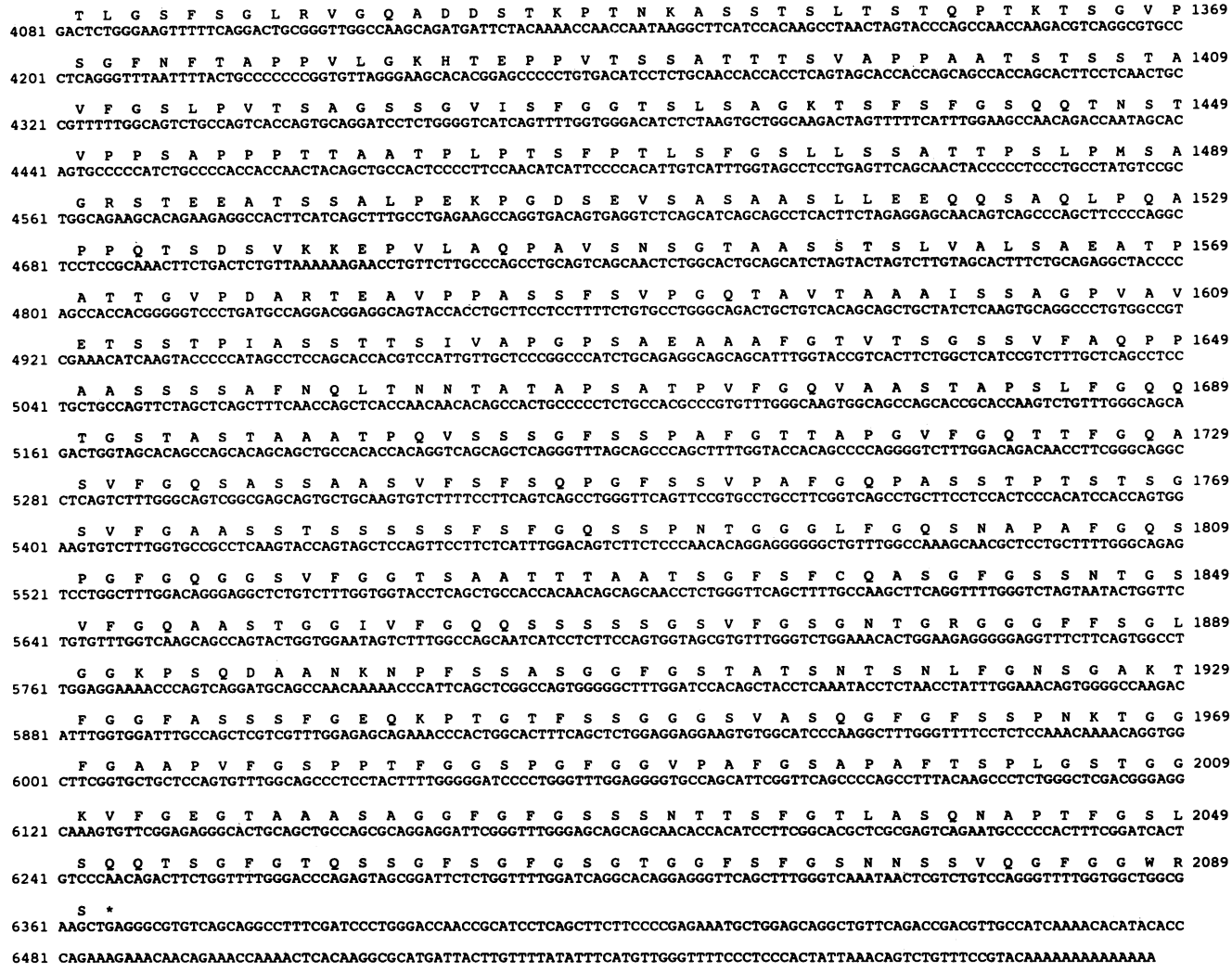

FIG. 2. cDNA sequence and putative amino acid sequence of the CAN protein. The position of the t(6;9) breakpoint is indicated by a solid triangle (nt 2530 to 2531). The putative leucine zipper (Fig. 6A, box A) and amphipathic helices (Fig. 6A, box B) are underlined.

tected an aberrant mRNA of 5.5 kb in the Northern blot of a sample from this patient (47). This result is in agreement with the observation that bone marrow cells from AUL patient SE contained a breakpoint in *icb*-9 but failed to show a breakpoint in band p23 on chromosome 6, which contains breakpoints of chromosomes from t(6;9) AML patients.

Cloning *dek* cDNA. To isolate a cDNA of the normal 2.7-kb transcript encoded on chromosome 6, a λgt11 cDNA library, derived from human testis RNA, was screened with probe DK2EX. In total, 24 clones were isolated and analyzed by restriction enzyme mapping. Two overlapping cDNA clones (DK9 and DK14) that contained 2.7 kb of contiguous sequences, probably representing the full-length transcript, were identified. The full-length *dek* cDNA clone of 2,699 nt was sequenced (Fig. 5). It contains an ORF of 1,125 nt encoding a putative protein of 375 amino acids and with an estimated molecular mass of 43 kDa, which is followed by a large 3′ untranslated region of 1,541 nt. The ATG codon is located at positions 34 to 36, which matches the Kozak consensus sequence (35). The predicted amino acid sequence of the ORF is shown in Fig. 5. The 3′ end of the cDNA sequence contains two AATAAA poly(A) addition sites next to each other, at positions 2682 and 2688, followed by a poly(A) stretch at position 2702.

As reported previously (47), the genomic area of chromosome 6p23, containing the breakpoints of chromosomes from four t(6;9) AML patients, was isolated in four overlapping λEMBL3 clones. The t(6;9) breakpoints appeared to map in a stretch of 12 kb in the middle of this region. The cDNA clones DK9 and DK14 were hybridized to Southern blots containing DNA of these genomic phages, which was digested with several restriction enzymes. Genomic fragments hybridizing to cDNA probes are present at either side of, but not within, the region that contains the chromosome 6 translocation breakpoints (Fig. 1B). Initially, the 12-kb breakpoint region was mapped by Southern blot analysis of genomic DNA and was delineated by two *Eco*RV sites. More precise mapping of the cloned chromosomal DNA reduced the size of the breakpoint region on chromosome 6p23 to an intron of 9 kb. Like what was done for the *can* gene, the intron-exon borders of the *dek* exons that flank

```
A: 5'can

atgcgaggtcaactgcgcgccgctcgggctcAGGGGAGGAAGTTTGCTGTCGAGCGGCCT   All3E4HP
                               AGGGGAGGAAGTTTGCTGTCGAGCGGCCT   hXT23

GGGTTCCGTGGGCAAGGCGTGGGAGGCAGCGTTGGCTCGTTC                     All3E4HP
GGGTTCCGTGGGCAAGGCGTGGGAGGCAGCGTTGGCTCGTTC                     hXT23

B: 3'can

ATATTTCATGTTGGGTTTTCCCTCCCACTATTAAACAGTCTGTTTCCGTACagaacgtat   All1F10.12St
ATATTTCATGTTGGGTTTTCCCTCCCACTATTAAACA                          hXT65
ATATTTCATGTTGGGTTTTCCCTCCCACTATTAAACAGTCTGTTTCCGTACAAAAAAAAA   hXE230

C: icb-9

GGATCCCAAGAGTGAAGCTCAGCTTCAGgtaggagatctatgtaaatctgtttaaaagatt   All4BE
GGATCCCAAGAGTGAAGCTCAGCTTCAGGAAATTCGGCGCCTTCATCAGTATGTGAAATTT   hXT37
tatttccctttaatttttttcttatagGAAATTCGGCGCCTTCATCAGTATGTGAAATTT    All7E2

D: 5'dek

cgccgcctcccagaacctcttcgtgccctcggcgtgccaGGCCCCGCGGCGGCCGAAATCC   MF3.20Hc1.4
                                       GGCCCGCGGCGGCCGAAATCC    DK14

GCGgtaaggggctgcttcgggacgcggagggcccggtcgcccctgccgcggccgacgggg    MF3.20Hc1.4
GCGGTTCACAGCATGTCCGCCTCGGCCCCTGCTGCGGAGGGGGAGGGAACCCCCACCCAG   DK14

E: 3'dek

TTATAAACCTTGTCAATAAAAATAAATCTAAATCactggtgttttaagtcacttgcattt   MF2.5E4.5
TTATAAACCTTGTCAATAAAAATAAATCTAAATCAAAAAAAAAAA                  DK9

gatatcttataggtgtatatagc                                        MF2.5E4.5

F: icb-6

CACAATGAAACAGATTTGCAAAAAGgtaattagacaaatgtttagattatttgctttgct   MF1.7E2.7
CACAATGAAACAGATTTGCAAAAAGGTCTATGAAAATTATCCTACTTATGATTTAACTGA   DK9
tatttttccttttcactatacatagGTCTATGAAAATTATCCTACTTATGATTTAACTGA   MF2.5E4.5
```

FIG. 3. Comparison of genomic and cDNA sequences of *dek* and *can*. (A) Genomic 700-bp *Hin*dIII-*Pst*I fragment (Al1F3E4HP) isolated from phage Al1F3 (Fig. 1A) contains the most 5′ sequences of cDNA clone hXT23. Presumably, these sequences belong to the first exon of *can*. (B) Genomic 900-bp *Stu*I fragment (Al1F10.12St) isolated from phage Al1F10.12 (Fig. 1A) contains the most 3′ exon of *can*. cDNA clone hXT65 ends immediately 3′ of the poly(A) signal; therefore, a 230-bp cDNA fragment (hXE230) that contains the poly(A) signal and poly(A) tail was generated by using the protocol for rapid amplification of cDNA ends. The poly(A) signal is underlined. (C) The intron-exon borders flanking *icb*-9 were determined to prove that the t(6;9) breakpoints are located in a single intron. A 1.2-kb *Bam*HI-*Eco*RI fragment (Al1F4BE) from phage Al1F4 (Fig. 1A) contains part of the exon flanking *icb*-9 upstream. In fact, the *Bam*HI site is located in the exon. A 2-kb *Eco*RI fragment (Al1F7E2) from phage Al1F7 (Fig. 1A) contains the exon flanking *icb*-9 downstream. The sequence of cDNA hXT37 shows both exons joined together. (D) A 1.4-kb *Hin*cII fragment (MF3.20Hc1.4) from the genomic *dek* phage MF3.20 (Fig. 1B) contains the most 5′ *dek* cDNA sequences of clone DK14 in a 240-bp *Bss*HII fragment. (E) *dek* has a large 3′ exon located in a 4.5-kb *Eco*RI fragment (MF2.5E4.5) of phage MF2.5 (Fig. 1B). The sequence encompassing the 3′ end of this exon is shown together with the 3′ end of cDNA clone DK9. (F) The intron-exon borders flanking *icb*-6 were determined. Downstream of *icb*-6 only one large 3′ exon is present in a 4.5-kb *Eco*RI fragment (MF2.5E4.5) of phage MF2.5 (Fig. 1B), the 5′ border of which is shown. The exon flanking *icb*-6 upstream is located in a 2.7-kb *Eco*RI fragment (MF1.7E2.7) of phage MF1.7 (Fig. 1B). The sequence of cDNA clone DK9 shows both exons joined together.

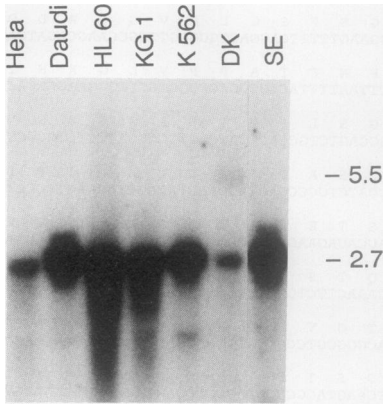

FIG. 4. Northern blot containing 20 μg of total RNA extracted from the cell lines HeLa, Daudi, HL60, KG1, and K562 and from bone marrow cells from t(6;9) AML patient DK and AUL patient SE. The sizes of the mRNAs hybridizing to the *dek* cDNA probe DK14 are indicated in kilobases.

icb-6 were sequenced. This showed that the translocation breakpoints occur between codons 349 and 350, almost at the C terminus of the *dek* ORF (Fig. 5 and 3F). The intron containing the breakpoints on chromosome 6 was termed *icb*-6. Fusion of *dek* and *can* via the introns *icb*-9 and *icb*-6 results in transcription of a chimeric mRNA in which the ORFs of *dek* and *can* are merged without disruption of their original reading frames. As a result, the DEK-CAN protein has a predicted molecular mass of 165 kDa.

Analysis of the *dek* gene. Hybridization of *dek* cDNA probes to the previously isolated λEMBL3 phages occurred 3′ of *icb*-6, to only one apparently continuous stretch of genomic DNA. As the restriction maps of this genomic DNA and the 3′ *dek* cDNA are colinear, it is likely that only a single 1.6-kb 3′ exon is present downstream of *icb*-6. To further substantiate this point, the 3′ end of this exon was sequenced and was found to be colinear with *dek* cDNA down to the poly(A) tail (Fig. 3E). However, a tiny intron in this exon can be excluded only by complete sequencing of the exon.

Hybridization of 5′ *dek* cDNA probes to genomic phages indicated that the 5′ end of the *dek* gene was not contained within them. Therefore, a genomic λEMBL3 library was screened with a 270-bp *Eco*RI-*Hin*cII fragment derived from the 5′ end of cDNA clone DK14. Phages (24) were isolated and characterized. Two overlapping phages that spanned the largest stretch of DNA were phage MF3.19 and MF3.20 (Fig. 1B). MF3.20 contains five *Bss*HII sites mapping close together in a 1.4-kb *Hin*cII fragment (MF3.20 Hc1.4). MF3.20Hc1.4 is the most 5′ fragment hybridizing to the 270-bp *Eco*RI-*Hin*cII cDNA probe. Fine mapping and subsequent sequence analysis showed that the most 5′ cDNA sequences are present in a 240-bp *Bss*HII fragment which is preceded by a region rich in G/C (Fig. 3D).

*Bss*HII sites are mostly clustered in CpG islands which often appear to represent promoter sequences (7). Therefore, this region in *dek* may well encompass the promoter area. Moreover, linking of a 520-bp genomic fragment, mapping immediately 5′ of the cDNA homologous sequence, to a chloramphenicol acetyltransferase reporter gene showed that this DNA region contains a strong promoter activity (results not shown).

The 3′ end of MF3.19 did not overlap with the most 5′ end of the previously isolated phages (MF1.2); a gap is still present in the map. Long-range mapping analysis (unpublished results) indicated that the distance between the translocation breakpoint of the chromosome from patient DK in *icb*-6 and the *Bss*HII sites is approximately 30 kb. As 27 kb of DNA between the *Bss*HII sites and the breakpoint of the chromosome from patient DK is present in the genomic phages, the gap is estimated to measure only a few kilobases.

Analysis of the *can* and *dek* cDNA sequences. Comparison of the predicted amino acid sequences of both *can* and *dek* with the EMBL data base failed to reveal any substantial

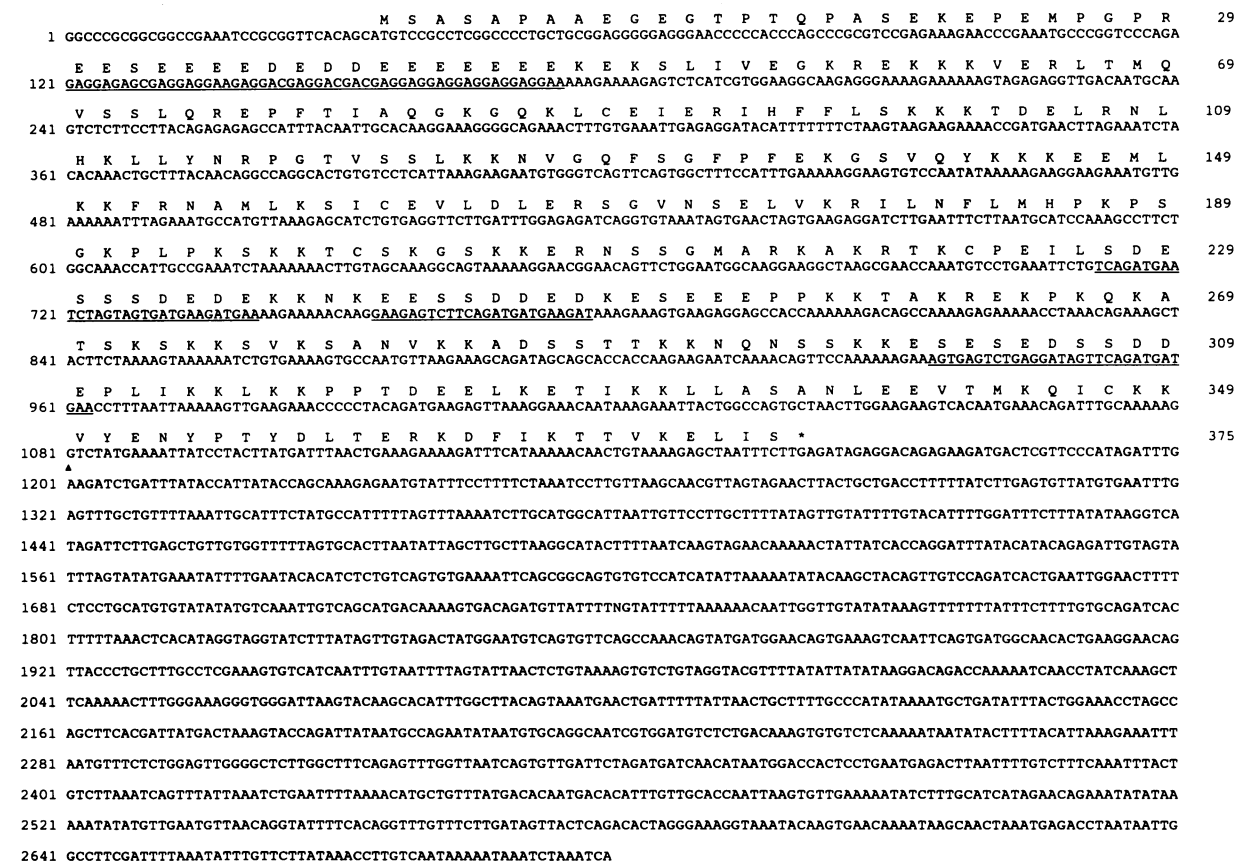

FIG. 5. cDNA sequence and putative amino acid sequence of the DEK protein. The position of the t(6;9) breakpoint is indicated with a solid triangle (nt 1080 to 1081). Acidic regions are underlined.

homology to known protein sequences. However, detailed analysis identified some structures that may have functional significance. In CAN, from amino acids (aa) 736 to 774, N terminal of *icb*-9, a leucine zipper motif is located. The leucine repeat consists of L-740, L-747, I-754, L-761, and L-768 and may represent a protein-protein dimerization domain. Projected on a helical wheel (Fig. 6A), hydrophobic residues at position +1 (relative to the leucines), acidic residues at position −1, and basic residues at position +2 are present. These residues may stabilize the formation of protein dimers through additional electrostatic interactions between the leucine repeats of proteins.

Immediately C terminal of the breakpoint in CAN from aa 811 to 887 two putative amphipathic helices are present and are separated by a loop of 25 residues, of which 13 are charged either positively (7) or negatively (6). The hydrophobic backbone of the first amphipathic helix is formed by I-814, L-817, V-821, A-824, V-828, and V-831 (Fig. 6A). The second amphipathic helix contains a heptad leucine repeat consisting of L-861, I-868, L-875, and L-882 (Fig. 6A). A region encompassing the C-terminal part of the loop and the C-terminal amphipathic helix (aa 840 to 887) shows homology to the human estrogen receptor (ER) dimerization domain: 30% of the residues is identical, 57% similar (Fig. 6B). The C-terminal 22 aa of this homology region in the ER were shown to be essential for the formation of ER homodimers (38).

Many SP or TP and SS, ST, TS, or TT dimers are present both N terminal and C terminal of the amino acid stretch containing the putative leucine zipper and amphipathic helices (Fig. 6A). This sequence motif has been proposed to have an ancillary role in DNA binding. At the C-terminus there is a recurrence of phenylalanine residues often in combination with S/T-P or S/T-S/T dimers.

In the predicted DEK protein no specific structures could be recognized apart from a continuous stretch of acidic residues at the N terminus, three acidic regions interspersed with serines, and a very high overall percentage (42%) of charged amino acids (H, R, K, E, and D).

Expression of *dek* and *can*. The expression patterns of *dek* and *can* in different mouse tissues may give a clue to the possible function of these genes. Twenty micrograms of total RNA of bone marrow, spleen, thymus, brain, liver, kidney, testes, ovary, placenta, and whole embryos 10, 13, 16, and 19 days after conception was loaded on a denaturing agarose gel. Hybridization of *dek* and *can* cDNA probes to hamster- and mouse-derived hybrid cell lines showed that both genes are conserved between species (unpublished results). Thus, the human *dek* cDNA clone DK14 and *can* cDNA clones hXT37 and hXT56 were used to screen for mouse *dek* and *can* transcripts. As shown in Fig. 7, *dek* and *can* are expressed in all tissues. *dek* is expressed at a relatively high level, while *can* seems to have a more restricted expression pattern. *can* expression was easily detected in RNA of thymus, spleen, bone marrow, kidney, brain, and testes but was hardly visible in all other tissues or in whole embryos during development.

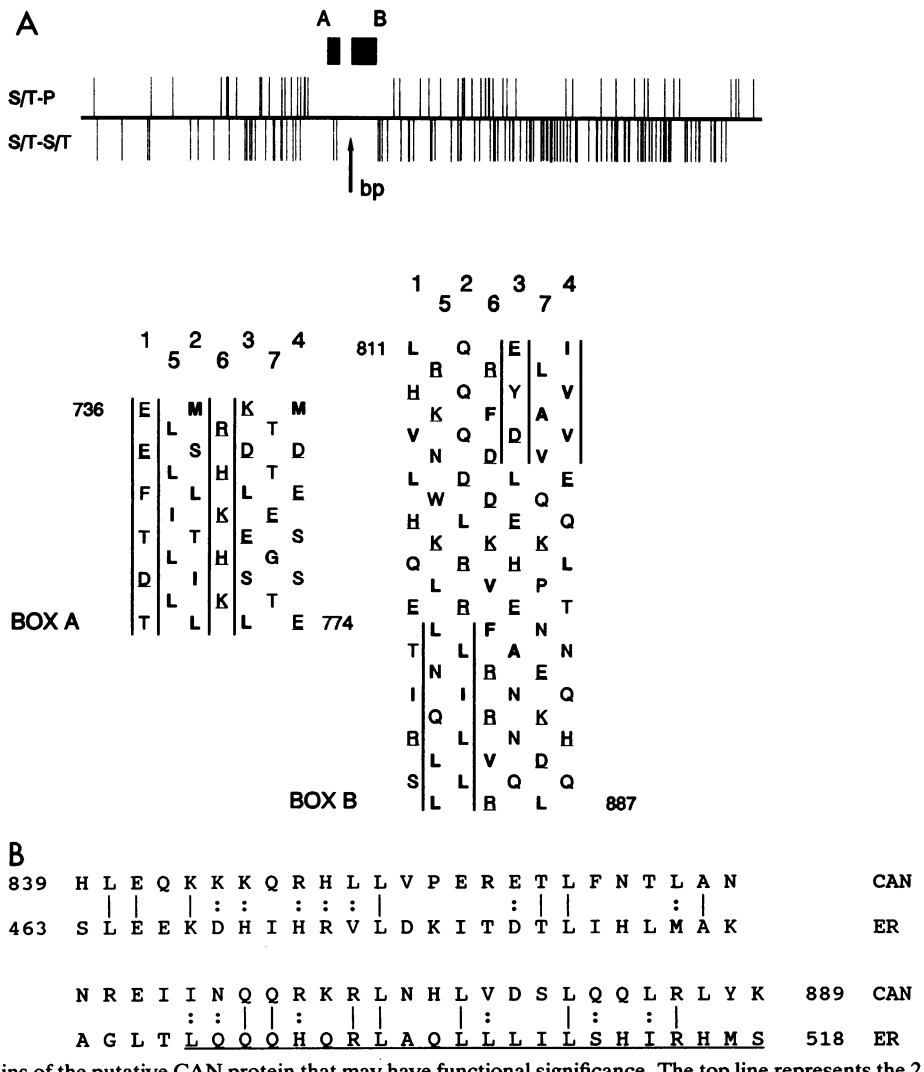

FIG. 6. (A) Domains of the putative CAN protein that may have functional significance. The top line represents the 2,192-aa putative CAN protein. The positions of S/T-P and S/T-S/T dimers are indicated with vertical lines. Box A (aa 736 to 775) represents a putative leucine zipper. Box B (aa 811 to 887) represents two amphipathic helices separated by a region of charged amino acids. An arrow indicates the position of the translocation breakpoint between boxes A and B. The amino acid sequences of boxes A and B are given underneath the CAN protein domains. The first four amino acid residues are written in a horizontal row and the next three are placed below and between them. In this way, the sequence can be read as an helical wheel, cut open at one side. Charged amino acids are underlined, and hydrophobic residues are in boldface type. Vertical lines indicate hydrophobic or charged sides of the predicted helical structure. (B) A part of the predicted CAN protein sequence present in box B (shown above) is homologous to the human ER. The homologous sequences are aligned. Identical amino acids are indicated by vertical lines, and similar amino acids are indicated by colons. The C-terminal 22 aa of the ER (underlined) are essential for ER homodimerization.

DISCUSSION

A novel fusion gene is present in leukemic cells carrying t(6;9) (p23;q34). The translocation breakpoints on chromosome 9 occur in one intron of the *can* gene, *icb*-9. Translocation breakpoints on chromosome 6 occur in one intron of the *dek* gene, *icb*-6. As a result of the translocation, a *dek-can* fusion gene encoding a chimeric *dek-can* transcript is generated. The sequence of this chimeric cDNA predicts it to encode a 165-kDa DEK-CAN protein.

Although the precise position of the breakpoints in *icb*-9 and *icb*-6 may vary, the same exons of *dek* and *can* are joined by splicing of the primary transcript of the fusion gene. The invariable *dek-can* transcript can be used as a marker of t(6;9) AML that can be sensitively monitored by the polymerase chain reaction (44). This may be a great advantage for diagnosis, monitoring of response to chemotherapy, and detection of minimal residual disease after bone marrow transplantation.

If steady-state levels of *dek-can* and *dek* transcripts in bone marrow from patient DK are compared, it appears that *dek* mRNA is much more abundant than *dek-can* mRNA. The bone marrow from patient DK contains >90% leukemic cells, of which every cell contains one chromosome 6 and one chromosome 6p−. In this cell population, the overall number of alleles of the normal *dek* gene and the fusion gene are about equal, and both are driven by the *dek* promoter. Higher steady-state levels of *dek* mRNA could be due to a longer half-life of *dek* transcripts compared with that of *dek-can* transcripts. Alternatively, enhancer sequences which are involved in transcription activation could be

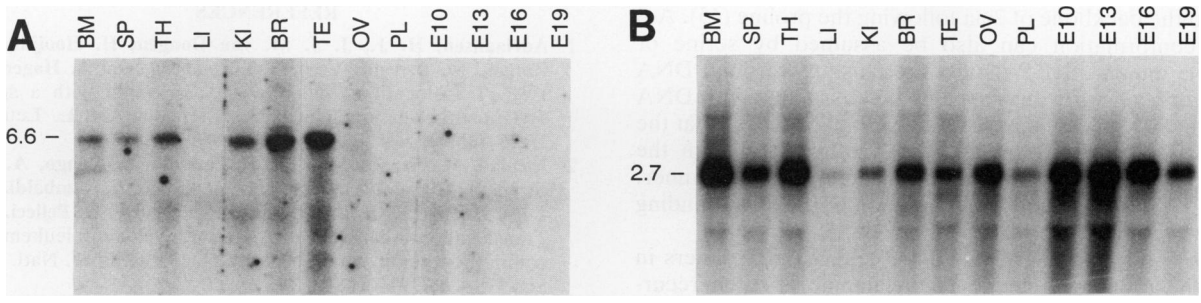

FIG. 7. Northern blot containing total RNA of various mouse tissues hybridized to *can* cDNA probe hXT37 and hXT56 (A) and to *dek* cDNA probe DK14 (B). The size of the transcript is indicated in kilobases. Abbreviations: BM, bone marrow; SP, spleen; TH, thymus; LI, liver; KI, kidney; BR, brain; TE, testes; OV, ovary; PL, placenta (13 days after conception); E10, E13, E16, and E19, embryos aged 10, 13, 16, and 19 days after conception, respectively.

present at the 3′ side of the *dek* gene. The enhancer would be removed from the fusion gene by the translocation.

The cellular function of DEK and CAN and the way DEK-CAN may interfere with normal hematopoiesis are still obscure. Neither of the two genes shows expression that is confined to the hematopoietic system. In fact, screening of a Northern blot containing RNA samples of different mouse tissues showed that *dek* is expressed ubiquitously. *can* is also expressed in all tissues, though at much lower and more variable levels. The tissues expressing *can* at a relatively high level include spleen and bone marrow. Since *can* mRNA is also found in human hematopoietic cell lines, it is unlikely that, because of the translocation, ectopic expression of *can* in hematopoietic cells would directly be involved in leukemogenic transformation. More likely, replacement of the N-terminal part of CAN by DEK sequences generates a protein that has different properties and is involved in transformation.

A breakpoint in *can* (*icb*-9) was also demonstrated in bone marrow cells from an AUL patient (SE) with an apparently normal karyotype (47). However, no breakpoint could be found in *dek*. In concordance with this observation, an aberrant transcript of 5.5 kb, detected by 3′ *can* probes in bone marrow RNA from this AUL patient, failed to hybridize to 5′ *dek* probes. cDNA cloning results strongly suggest that in this patient, *can* forms a fusion gene in which the 5′ sequences are derived from another, as yet unknown gene (46a). Therefore, it is possible that the C-terminal part of CAN contains domains involved in the leukemogenic process that may be activated by different N-terminal moieties.

Preliminary protein localization data were obtained by immunocytochemistry with antibodies directed against DEK and CAN and COS cells transiently expressing high levels of CAN, DEK, or DEK-CAN protein. CAN appears to be mainly cytoplasmic, while DEK has a strictly nuclear localization. The fusion of DEK to CAN results in a protein with a nuclear localization (19a).

In view of these data, the analysis of the *can* cDNA sequence revealed some structures that may be indicative for its function.

(i) An amphipathic helix with a heptad leucine repeat is predicted by the sequence just 5′ of *icb*-9. This leucine zipper motif has been detected in many proteins such as FOS, JUN, GCN4, and CCAAT/enhancer binding protein (10, 50), in which it mediates the formation of either homo- or heterodimers. A basic stretch of amino acids, juxtaposed to the leucine zipper, can function as a DNA binding element. In CAN no basic region is present adjoining the leucine zipper, and hence this helix most likely functions as a dimerization domain. Not only the addition of novel sequences to the 3′ part of *can* but also the removal of the original 5′ part of the gene may contribute to the putatively tumorigenic properties of the *dek-can* fusion gene. As the leucine zipper is detached from the C-terminal CAN sequences by the translocation, it is tempting to speculate that this structure may be the interaction site for a factor that could regulate CAN activity.

(ii) The protein sequence just 3′ of *icb*-9 predicts two amphipathic helices separated by a stretch of 25 amino acids, containing many charged residues. Several arguments suggest that this domain may function in protein dimerization. (a) The C-terminal putative helix and part of the preceding charged amino acids show homology to the hormone binding region of the human and mouse ERs (23). It has been shown that the mouse ER contains a strong dimerization domain adjoining the hormone binding domain (18, 38). The entire sequence containing both domains is conserved within the steroid receptor family. At the N-terminal side of this domain, the homology between the ER and CAN extends beyond the homology between the ER and other steroid hormone receptors. However, CAN has no homology to the hormone binding domain immediately C-terminal of the dimerization domain. It is interesting that the homologous protein domain in another member of the steroid hormone receptor family, the retinoic acid receptor type α, was shown to dimerize with multiple cell-type-specific proteins which have not yet been characterized. Dimerization increased the affinity of the receptor for its cognate binding sequence (22). In addition, homology of CAN to the ER is noteworthy with regard to the finding that the retinoic acid receptor type α is involved in t(15;17) in acute promyelocytic leukemia (2, 8, 15). It will be interesting to analyze whether CAN can form heterodimers with the ER or other members of the steroid hormone receptor family. (b) Although no direct homology is present, the putative structure of CAN just C-terminal of *icb*-9 (aa 811 to 887) architecturally resembles aa 82 to 162 of transcription factor AP-4, a basic stretch-helix-loop-helix protein (30). This part of AP-4 contains an additional dimerization domain, which, like this region in CAN, consists of two amphipathic helices separated by a stretch of 28 aa, containing many charged residues (30).

(iii) Many SP and TP dimers are present both N-terminal and C-terminal of the region containing the putative leucine zipper and amphipathic helices. A proline preceded by a serine or a threonine forms a β turn I, which is stabilized by formation of hydrogen bonds between the serine or threo-

nine and the backbone of 2 aa following the proline (45). A β turn I conformation can also be assumed by serine or threonine dimers. S/T-P dimers are clustered around DNA binding domains of many proteins that associate with DNA in a sequence-specific manner. Suzuki (45) proposes that the S/T-P-X-X (X for any amino acid) motif will bind in the minor groove of DNA in a sequence-independent manner. This may stabilize a specific interaction of the DNA binding motif in the major groove.

In the C-terminal cluster of S/T-P and S/T-S/T dimers in CAN, the aromatic residue phenylalanine is often recurring. The C-terminal part of RNA polymerase II of both the yeast *Saccharomyces cerevisiae* and mammals contains a SPTSPSY repeat (3, 13), which is essential for its function (4). Suzuki (46) argues that the β turn I-X-Y motif may be essential for DNA binding and shows that the aromatic ring of the tyrosine residue in this repeat can intercalate into the DNA. In *Drosophila* RNA polymerase II, tyrosine is replaced by another aromatic residue, phenylalanine (6). A structure of β turns combined with aromatic residues is therefore postulated to be a novel type of DNA binding domain. In the 3′ part of CAN a S/T-S/T/P-X-F sequence occurs 14 times. We will study whether this region has DNA binding capacity, either by itself or by stabilizing DNA binding domains of transcription factor complexes that contain the CAN protein.

The predicted protein sequence of DEK contains a remarkably high percentage of charged amino acids. At the N terminus (aa 30 to 47), DEK contains a continuous stretch of acidic residues. Three other acidic stretches are present, from aa 227 to 236, 241 to 248, and 301 to 310. They contain acidic residues interspersed only by serine residues. Acidic regions were found mainly in two types of nuclear proteins (17). (i) Chromatin-associated proteins such as nucleolin and high mobility group proteins contain acidic regions that can interact with the basic domains of histones (36, 37). These proteins also contain a conserved DNA binding domain, the high mobility group box, a sequence motif that is not present in DEK. (ii) A class of transcriptional activators, among which are herpes simplex virus VP16 protein and the yeast transcription factor GCN4, contain an acidic patch that can interact with the RNA polymerase II complex (9, 29).

Many basic amino acids are present in the DEK protein next to the acidic regions. The calculated pI of DEK is 8.9. Because of these basic stretches, several putative nuclear localization signals can be recognized. DEK is completely devoid of hydrophobic stretches.

We speculate that replacement of N-terminal CAN sequences by almost the entire DEK protein may activate the transforming potential of CAN. However, the mechanism of this putative activation remains to be determined. Analysis of the primary structure of DEK and CAN combined with the preliminary localization data suggests that these proteins may have a function in the cell nucleus.

Up to now, breakpoints of three different translocations in myeloid leukemia have been cloned and molecularly analyzed. Thus far the formation of fusion genes seems to be the predominant effect of translocations in myeloid leukemia.

ACKNOWLEDGMENTS

We thank D. Bootsma and D. Meijer for continuous support and discussion; J. Abels, H. Adriaansen, A. Hagemeijer, and B. Löwenberg for patient material; and M. Kuit and T. de Vries Lentsch for photographic work. We are indebted to A. Coulson (MRC, Edinburgh) for help with computer homology searches.

This work was supported by the Dutch Cancer Foundation.

REFERENCES

1. **Adriaansen, H. J., J. J. M. van Dongen, H. Hooijkaas, K. Hählen, M. B. van't Veer, B. Löwenberg, and A. Hagemeijer.** 1988. Translocation (6;9) may be associated with a specific TdT-positive immunological phenotype in ANLL. Leukemia **2**:136–140.
2. **Alcalay, M., D. Zangrilli, P. P. Pandolfi, L. Longo, A. Mencarelli, A. Giacomucci, M. Rocchi, A. Biondi, A. Rambaldi, F. Lo Coco, D. Diverio, E. Donti, F. Griniani, and P. G. Pelicci.** 1991. Translocation breakpoint of acute promyelocytic leukemia lies within the retinoic acid receptor alpha locus. Proc. Natl. Acad. Sci. USA **88**:1977–1981.
3. **Allison, L. A., M. Moyle, M. Shales, and C. J. Ingles.** 1985. Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases. Cell **42**:599–610.
4. **Allison, L. A., J. K. Wong, V. D. Fitzpatrick, M. Moyle, and C. J. Ingles.** 1988. The C-terminal domain of the largest subunit of RNA polymerase II of *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and mammals: a conserved structure with an essential function. Mol. Cell. Biol. **8**:321–329.
5. **Auffray, C., and F. Rougeon.** 1980. Purification of mouse immunoglobulin heavy-chain messenger RNAs from total myeloma tumor RNA. Eur. J. Biochem. **107**:303–314.
6. **Biggs, J., L. L. Searles, and A. L. Greenleaf.** 1985. Structure of the eukaryotic transcription apparatus: features of the gene for the largest subunit of Drosophila RNA polymerase II. Cell **42**:611–621.
7. **Bird, A. P.** 1986. CpG islands and the function of DNA methylation. Nature (London) **321**:209–213.
8. **Borrow, J., A. D. Goddard, D. Sheer, and E. Solomon.** 1990. Molecular analysis of acute promyelocytic leukemia breakpoint cluster region on chromosome 17. Science **249**:1577–1580.
9. **Brent, R., and M. Ptashne.** 1985. A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor. Cell **43**:729–736.
10. **Busch, S. J., and P. Sassone-Corsi.** 1990. Dimers, leucine zippers and DNA-binding domains. Trends Genet. **6**:36–40.
11. **Chirgwin, J. M., A. E. Przybyla, J. MacDonald, and W. J. Rutter.** 1979. Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. Biochemistry **18**:5294–5299.
12. **Chomienne, C., P. Ballerini, N. Balitrand, M. E. Huang, I. Krawice, S. Castaigne, P. Fenaux, P. Tiollais, A. Dejean, L. Degos, and H. de Thé.** 1990. The retinoic acid receptor alpha gene is rearranged in retinoic acid-sensitive promyelocytic leukemias. Leukemia **4**:802–807.
13. **Corden, J. L., D. L. Cadena, J. J. Ahearn, and M. E. Dahmus.** 1985. A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II. Proc. Natl. Acad. Sci. USA **82**:7934–7938.
14. **Daley, G. Q., R. Van Etten, and D. Baltimore.** 1990. Induction of chronic myelogenous leukemia in mice by the P210bcr/abl gene of the Philadelphia chromosome. Science **247**:824–830.
15. **de Thé, H., C. Chomienne, M. Lanotte, L. Degos, and A. Dejean.** 1990. The t(15;17) translocation of acute promyelocytic leukaemia fuses the retinoic acid receptor alpha gene to a novel transcribed locus. Nature (London) **347**:558–561.
16. **De Thé, H., C. Lavau, A. Marchio, C. Chomienne, L. Degos, and A. Dejean.** 1991. The PML-RARα fusion mRNA generated by the t(15;17) translocation in acute promyelocytic leukemia encodes a functionally altered RARα. Cell **66**:675–684.
17. **Earnshaw, W. C.** 1987. Anionic regions in nuclear proteins. J. Cell Biol. **105**:1479–1482.
18. **Fawell, S. E., J. A. Lees, R. White, and M. G. Parker.** 1990. Characterization and colocalization of steroid binding and dimerization activities in the mouse estrogen receptor. Cell **60**:953–962.
19. **Feinberg, A. P., and B. Vogelstein.** 1983. A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. **132**:6–13.

19a. **Fornerod, M., M. von Lindern, and G. Grosveld.** Unpublished data.

20. **Fourney, R. M., J. Miyakoshi, R. S. Day III, and M. C. Paterson.**

1988. Northern blotting: efficient RNA staining and transfer. Focus **10**:5–6.

21. **Frohman, M. A., M. K. Dash, and G. R. Martin.** 1988. Rapid production of full-length cDNAs from rare transcripts amplification using a single gene-specific oligonucleotide primer. Proc. Natl. Acad. Sci. USA **85**:8998–9002.
22. **Glass, C. K., O. V. Devary, and M. G. Rosenfeld.** 1990. Multiple cell type-specific proteins differentially regulate target sequence recognition by the alpha retinoic acid receptor. Cell **63**:729–738.
23. **Green, S., P. Walter, V. Kumar, A. Krust, J. M. Bornert, P. Argos, and P. Chambon.** 1986. Human oestrogen receptor cDNA: sequence, expression and homology to v-erb-A. Nature (London) **320**:134–139.
24. **Gubler, U., and B. J. Hoffman.** 1983. A simple and very efficient method for generating cDNA libraries. Gene **25**:263–269.
25. **Heisterkamp, N., G. Jenster, H. J. Ten Hoeve, D. Zovich, P. K. Pattengale, and J. Groffen.** 1990. Acute leukaemia in bcr/abl transgenic mice. Nature (London) **344**:251–253.
26. **Hermans, A., J. Gow, L. Selleri, M. von Lindern, A. Hagemeijer, L. M. Wiedemann, and G. Grosveld.** 1988. bcr-abl oncogene activation in Philadelphia chromosome-positive acute lymphoblastic leukemia. Leukemia **2**:628–633.
27. **Hermans, A., N. Heisterkamp, M. von Lindern, S. van Baal, D. Meijer, D. van der Plas, L. M. Wiedemann, J. Groffen, D. Bootsma, and G. Grosveld.** 1987. Unique fusion of bcr and c-abl genes in Philadelphia chromosome positive acute lymphoblastic leukemia. Cell **51**:33–40.
28. **Hermans, A., L. Selleri, J. Gow, L. Wiedeman, and G. Grosveld.** 1989. Molecular analysis of the Philadelphia translocation in myelogenous and acute lymphoblastic leukemia. Cancer Cells **7**:21–26.
29. **Hope, I. A., and K. Struhl.** 1986. Functional dissection of a eukaryotic transcriptional activator protein, GCN4 of yeast. Cell **46**:885–894.
30. **Hu, Y. F., B. Luscher, A. Admon, N. Mermod, and R. Tjian.** 1990. Transcription factor AP-4 contains multiple dimerization domains that regulate dimer specificity. Genes Dev. **4**:1741–1752.
31. **Huynh, T. V., R. A. Young, and R. W. Davis.** 1984. Construction and screening cDNA libraries in λgt10 and λgt11, p. 49–78. *In* D. Glover (ed.), DNA cloning techniques: a practical approach. IRL Press, Oxford.
32. **Kadowaki, T., H. Kadowaki, and S. I. Tayler.** 1990. A nonsense mutation causing decreased levels of insulin receptor mRNA: detection by a simplified technique for direct sequencing of genomic DNA amplified by the polymerase chain reaction. Proc. Natl. Acad. Sci. USA **87**:658–662.
33. **Kakizuka, A., W. H. Miller, K. Umesono, R. P. Warrell, S. R. Frankel, V. V. V. S. Murty, E. Dmitrovsky, and R. M. Evans.** 1991. Chromosomal translocation t(15;17) in human acute promyelocytic leukemia fuses RARα with a novel putative transcription factor, PML. Cell **66**:663–674.
34. **Konopka, J. B., S. M. Watanabe, and O. N. Witte.** 1984. An alteration of the human c-abl protein in K562 leukemia cells unmasks associated tyrosine kinase activity. Cell **37**:1035–1042.
35. **Kozak, M.** 1988. A profusion of controls. J. Cell Biol. **107**:1–7.
36. **Krohne, G., and W. W. Franke.** 1980. A major soluble acidic protein located in nuclei of diverse vertebrate species. Exp. Cell Res. **129**:167–189.
37. **Lapeyre, B., H. Bourbon, and F. Amalric.** 1987. Nucleolin, the major nucleolar protein of growing eukaryotic cells: an unusual protein structure revealed by the nucleotide sequence. Proc. Natl. Acad. Sci. USA **84**:1472–1476.
38. **Lees, J. A., S. E. Fawell, R. White, and M. G. Parker.** 1990. A 22-amino-acid peptide restores DNA-binding activity to dimerization-defective mutants of the estrogen receptor. Mol. Cell. Biol. **10**:5529–5531.
39. **Pearson, M. G., J. W. Vardiman, B. M. LeBean, J. D. Rowley, S. Schwartz, S. L. Kerman, M. M. Cohen, E. W. Fleischman, and E. L. Prigogina.** 1985. Increased numbers of marrow basophils may be associated with a t(6;9) in ANLL. Am. J. Hematol. **18**:393–403.
40. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
41. **Sandberg, A. A., R. Morgan, J. A. McCallister, M. B. Kaiser, and F. Hecht.** 1983. Acute myeloblastic leukemia (AML) with t(6;9) (p23;q34): a specific subgroup of AML? Cancer Genet. Cytogenet. **10**:139–142.
42. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
43. **Shtivelman, E., B. Lifshitz, R. P. Gale, and E. Canaani.** 1985. Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. Nature (London) **315**:550–554.
44. **Soekarman, D., M. von Lindern, S. Daenen, B. de Jong, C. Fonatsch, B. Heinze, C. R. Bartram, A. Hagemeijer, and G. Grosveld.** Submitted for publication.
45. **Suzuki, M.** 1989. SPXX, a frequent sequence motif in gene regulatory proteins. J. Mol. Biol. **207**:61–84.
46. **Suzuki, M.** 1990. The heptad repeat in the largest subunit of RNA polymerase II binds by intercalating into DNA. Nature (London) **344**:562–565.

46a. **von Lindern, M.** Unpublished data.

47. **von Lindern, M., A. Poustka, H. Lerach, and G. Grosveld.** 1990. The (6;9) chromosome translocation, associated with a specific subtype of acute nonlymphocytic leukemia, leads to aberrant transcription of a target gene on 9q34. Mol. Cell. Biol. **10**:4016–4026.
48. **von Lindern, M., T. van Agthoven, A. Hagemeijer, H. Adriaansen, and G. Grosveld.** 1989. The human pim-1 gene is not directly activated by the translocation (6;9) in acute nonlymphocytic leukemia. Oncogene **4**:75–79.
49. **Walker, L. C., T. S. Ganesan, S. Dhut, B. Gibbons, T. A. Lister, J. Rothbard, and B. D. Young.** 1987. Novel chimaeric protein expressed in Philadelphia positive acute lymphoblastic leukaemia. Nature (London) **329**:851–853.
50. **Ziff, E. B.** 1990. Transcription factors: a new family gathers at the cAMP response site. Trends Genet. **6**:69–72.