ELSEVIER

Gene 262 (2001) 23–33

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

# Architecture and anatomy of the genomic locus encoding the human leukemia-associated transcription factor RUNX1/AML1 ☆

Ditsa Levanon[a,1], Gustavo Glusman[a,1], Thorsten Bangsow[a,2], Edna Ben-Asher[a], Dean A. Male[a,3], Nili Avidan[a], Carmen Bangsow[a], Masahira Hattori[b], Todd D. Taylor[b], Stefan Taudien[c], Karin Blechschmidt[c], Nobuyoshi Shimizu[d], Andre Rosenthal[c], Yoshiyuki Sakaki[b], Doron Lancet[a], Yoram Groner[a,*]

[a]*Dept of Molecular Genetics and Human Genome Center, The Weizmann Institute of Science, Rehovot, 76100, Israel*
[b]*RIKEN, Genomic Sciences Center, Sagamihara 228-8555, Japan*
[c]*Institut für Molekulare Biotechnologie, Genomanalyse, D-07745, Jena, Germany*
[d]*Dept. of Molecular Biology, Keio University School of Medicine, Tokyo 160-8582, Japan*

Received 16 September 2000; accepted 2 November 2000
Received by A.J. van Wijnen

**Abstract**

The *RUNX1* gene on human chromosome 21q22.12 belongs to the 'runt domain' gene family of transcription factors (also known as AML/CBFA/PEBP2α). *RUNX1* is a key regulator of hematopoiesis and a frequent target of leukemia associated chromosomal translocations. Here we present a detailed analysis of the *RUNX1* locus based on its complete genomic sequence. *RUNX1* spans 260 kb and its expression is regulated through two distinct promoter regions, that are 160 kb apart. A very large CpG island complex marks the proximal promoter (promoter-2), and an additional CpG island is located at the 3′ end of the gene. Hitherto, 12 different alternatively spliced *RUNX1* cDNAs have been identified. Genomic sequence analysis of intron/exon boundaries of these cDNAs has shown that all consist of properly spliced authentic coding regions. This indicates that the large repertoire of RUNX1 proteins, ranging in size between 20–52 kDa, are generated through usage of alternatively spliced exons some of which contain in frame stop codons. The gene's introns are largely depleted of repetitive sequences, especially of the LINE1 family. The *RUNX1* locus marks the transition from a ~1 Mb of gene-poor region containing only pseudogenes, to a gene-rich region containing several functional genes. A search for *RUNX1* sequences that may be involved in the high frequency of chromosomal translocations revealed that a 555 bp long segment originating in chromosome 11 *FLI1* gene was transposed into *RUNX1* intron 4.1. This intron harbors the t(8;21) and t(3;21) chromosomal breakpoints involved in acute myeloid leukemia. Interestingly, the *FLI1* homologous sequence contains a breakpoint of the t(11;22) translocation associated with Ewing's tumors, and may have a similar function in *RUNX1*. 

*Keywords*: Gene structure; Chromosomal translocations; FLI1 homology; RUNX family

## 1. Introduction

The Runt-related transcription factor 1 (RUNX1) (previously called AML1/CBFA2/PEBP2αB) belongs to a small family of transcription factors. Members of this family share homology in a 128 amino acids region, designated 'runt domain' (RD), first identified in the *Drosophila* pair-rule gene *runt*. This domain directs the binding of RUNX1 to the consensus core sequence YGYGGT of target genes and mediates RUNX1 interactions with its β-subunit, called core-binding factor β (CBFβ) (rev. in Ito and Bae, 1997; Downing, 1999). NMR spectroscopy and crystal structure analysis of the RD showed that it has an S-type immunoglobulin fold, establishing a structural relationship between the RD and DNA binding domains of NF-κB, NFAT1, p53 and

Abbreviations: All, acute lymphoid leukemia; AML, acute myeloid leukemia; CBFβ, core binding factor β; DS, Down Syndrome; RD, runt domain; TAD, transactivation domain; UTR, untranslated region; IRES, internal ribosomal entry site

☆ The nomenclature committee of the Human Genome Organization has recently adopted the following symbols to designate the genes for runt-related transcription factors: RUNX1 (alias AML1/CBFA2/PEBP2αB), *RUNX2* (alias AML3/CBFA1/PEBP2αA) and *RUNX3* (alias AML2/CBFA3/PEBP2αC).

* Corresponding author. Tel.: +972-8-934-3972; fax: +972-8-934-4108.
*E-mail address:* yoram.groner@weizmann.ac.il (Y. Groner).

[1] These authors contributed equally to this work.

[2] Present address: Klinikum der Johann Wolfgang Goethe-Universitaet, Zentrum fuer Innere Medizin - Molekulare Haematologie, Theodor-Stern-Kai 7, D-60590 Frankfurt am Main, Germany.

[3] Present address: Department of Microbiology and Infectious Diseases, Flinders University of South Australia, Bedford Park, S.A. 5042, Australia.

0378-1119/01/$ - see front matter 
PII: S0378-1119(00)00532-1

the STAT proteins (rev. in Ito, 1999; Warren et al., 2000). Three *RUNX* genes were identified in human and mice: *RUNX1/AML1* on human chromosome 21q22.12, *RUNX2/AML3* on human chromosome 6p21 and *RUNX3/AML2* on human chromosome 1p36 (Levanon et al., 1994; Ito and Bae, 1997). In adults, *RUNX 1* and *RUNX3* are expressed mainly in the hematopoietic system (Levanon et al., 1994; Satake et al., 1995; Levanon et al., 1996; Meyers et al., 1996). During embryonic development RUNX2 functions as a key regulator of osteogenesis (rev. in Ito, 1999; Speck et al., 1999) and RUNX1 plays an important role in the commitment of the hemangiogenic endothelium to produce definitive hematopoietic cells (Speck et al., 1999). This early hematopoietic expression of RUNX1 correlates with the lack of definitive hematopoiesis in homozygous *Runx1* knock-out mice (rev. in Speck et al., 1999). Human *RUNX1* and *CBFβ* are the most frequently targeted genes in chromosomal translocations associated with acute myeloid leukemia (AML) and acute lymphoid leukemia (ALL) (Look, 1997). The molecular mechanisms underlying this high frequency of leukemia-associated translocations are not known. *RUNX1* is truncated in the t(8;21) translocation which occurs in about 12% of AML-M2 patients and in the t(12;21) translocation occuring in 20% of patients with pro-B-cell ALL. The t(8;21) translocation creates a fused protein which contains the entire RD. This protein binds to DNA strongly and out-compete RUNX dependent transcription in vitro (Meyers et al., 1996). Chromosomal translocations represent one way by which perturbation in *RUNX1* function causes leukemia, but other alterations in its activity may prove to be leukemogenic as well. Recently, it has been reported that haploinsufficiency of *RUNX1* causes familial thrombocytopenia with propensity to develop AML (Song et al., 1999). Increased gene dosage of *RUNX1* occurs in Down Syndrome (DS), the phenotypic manifestation of trisomy 21, and DS patients have an increased risk to develop acute megakaryoblastic leukemia (AML-M7).

*RUNX1* exhibits a complex pattern of regulated expression, at the levels of transcription, splicing and translation (Miyoshi et al., 1995; Ghozi et al., 1996; Levanon et al., 1996; Pozner et al., 2000). Transcription of *RUNX1* is initiated at two distinct 5′ regions, a distal region-promoter-1 (P1) and a proximal region- promoter-2 (P2). These two promoters generate a large number of alternatively spliced mRNAs, that differ in their types of 5′ and 3′ UTRs and in their coding regions (Miyoshi et al., 1995; Levanon et al., 1996; Zhang et al., 1997). The full length coding regions harbor the carboxy-terminal half of the protein which regulate transcription (including the transactivation domain-TAD) (rev. in Downing, 1999; Ito, 1999). Other RUNX1 isoforms are shorter, lack TAD and have altered biological activities (Tanaka et al., 1995; Ben Aziz-Aloya et al., 1998; Downing, 1999). In vitro studies demonstrated that RUNX1 functions as an organizer of transcriptional active complexes, regulating the activity of several hematopoietic genes such as *TCRα*, *TCRβ*, *NP-3* and *CSF-1R* (Ito and Bae, 1997; Downing, 1999). Interestingly, RUNX1 can either activate or repress transcription of target genes, depending on the protein isoform studied and its ability to interact with other transcriptional regulators (rev. in Fisher and Caudy, 1998; Levanon et al., 1998; Downing, 1999).

The sequence of the entire *RUNX1* gene was recently established in the framework of the chromosome 21 sequencing consortium. Here we report the detailed analysis of the 260 kb sequence of the gene, the precise organization of all the alternatively spliced mRNA isoforms and several biologically significant structural features of the gene and its genomic locus.

## 2. Materials and methods

### 2.1. Computer analysis

#### 2.1.1. Sequence analysis

Sequence data were analyzed using the GESTALT Workbench (Glusman and Lancet, 2000) for genomic sequence visualization, as well as the comprehensive tool RUMMAGE-DP at http://gen100.imb-jena.de/rummage, which combines more than 25 different programs.

#### 2.1.2. Regional analysis

A 1.87 Mb contig was built by assembling GenBank entries AJ229041-043, AP000119-125 and AF027153. The sequence was split into non-overlapping 20 kb segments, and analyzed for the content of G + C, CpG islands (regions of at least 200 bp, with CpG CV $\geqq$ 0.6 and G + C content $\geqq$ 50%), repeats (using RepeatMasker version 290499) and genes (using GenScan and fgenes 1.6). Potential homologies were detected by blastx using predicted exons as queries, as well as all the repeat-masked segments as queries for blastn, FASTA, blastx and FASTX versus the non-redundant GenBank release 113.

#### 2.1.3. Analysis of introns

Introns were extracted from the primate partition of GenBank release 113, from entries at least 25,000 bp long, representing genomic clones. Introns were defined as any DNA sequence between contiguous exons in an annotated multi-exon coding sequence (GenBank tag: CDS). Each extracted intron of length $\geqq$ 500 bp was analyzed using RepeatMasker as described above. The dataset consisted of 8522 introns, totalling 34.89 Mb of sequence. The 11 introns of the *RUNX1* gene were similarly analyzed. For each *RUNX1* intron a 'similar length' set was defined, including those database introns of length between half and double that of the *RUNX1* intron.

### 2.2. Isolation of the RUNX1/FLI1 homologous region and Southern blot analysis

The 350 bp intronic fragment of the *RUNX1* gene that

shares a high degree of similarity with a sequence in the *FLI1* gene was amplified by PCR using the *RUNX1* specific YACs 72H9 and 860G11 obtained from Dr. D. Le Paslier, Centre d'Etude du Polymorphisme Humain (CEPH), Paris, France. The two oligonucleotides used for PCR were: 5′-TAAAAGTGAAAGAGCTGGCTG-3′; and 5′-GCTGCA-CATTTTACCTTACTC-3′. The 350 bp PCR product was cloned into the pGEM-T vector (Promega) and sequenced. Southern blots of human placental DNA were analyzed as described (Levanon et al., 1994).

## 3. Results and discussion

### 3.1. The structure of the RUNX1 gene

The complete sequence of the *RUNX1* gene was established within the framework of chromosome 21 mapping and sequencing project (Hattori et al., 2000). The overall structure of *RUNX1* was generated by a comparison between genomic and cDNA sequences (Fig. 1). The three *RUNX* genes are highly conserved in their structure and sequence. *RUNX3* contains the smallest number of exons all of which are conserved in the two other *RUNX* genes (Bangsow et al., manuscript in preparation). For this reason it was decided that *RUNX3* will serve as the structural prototype of the *RUNX* gene family and the numbering system of exons will be based on this gene (van Wijnen et al., manuscript in preparation). *RUNX1* spans 260 kb and contains 12 exons (Fig.1). Common exons to the three *RUNX* genes are numbered 1–6 while additional exons not present in *RUNX3* are indicated by subnumbers. All the cDNAs presented in Fig. 1 except for one (*RUNX1*/p51) are composed of properly spliced exons, indicating that the entire collection are bona fide *RUNX1* mRNAs. In addition to the two distinct 5′UTRs (Fig. 1), these cDNAs differ in their 3′UTRs as well

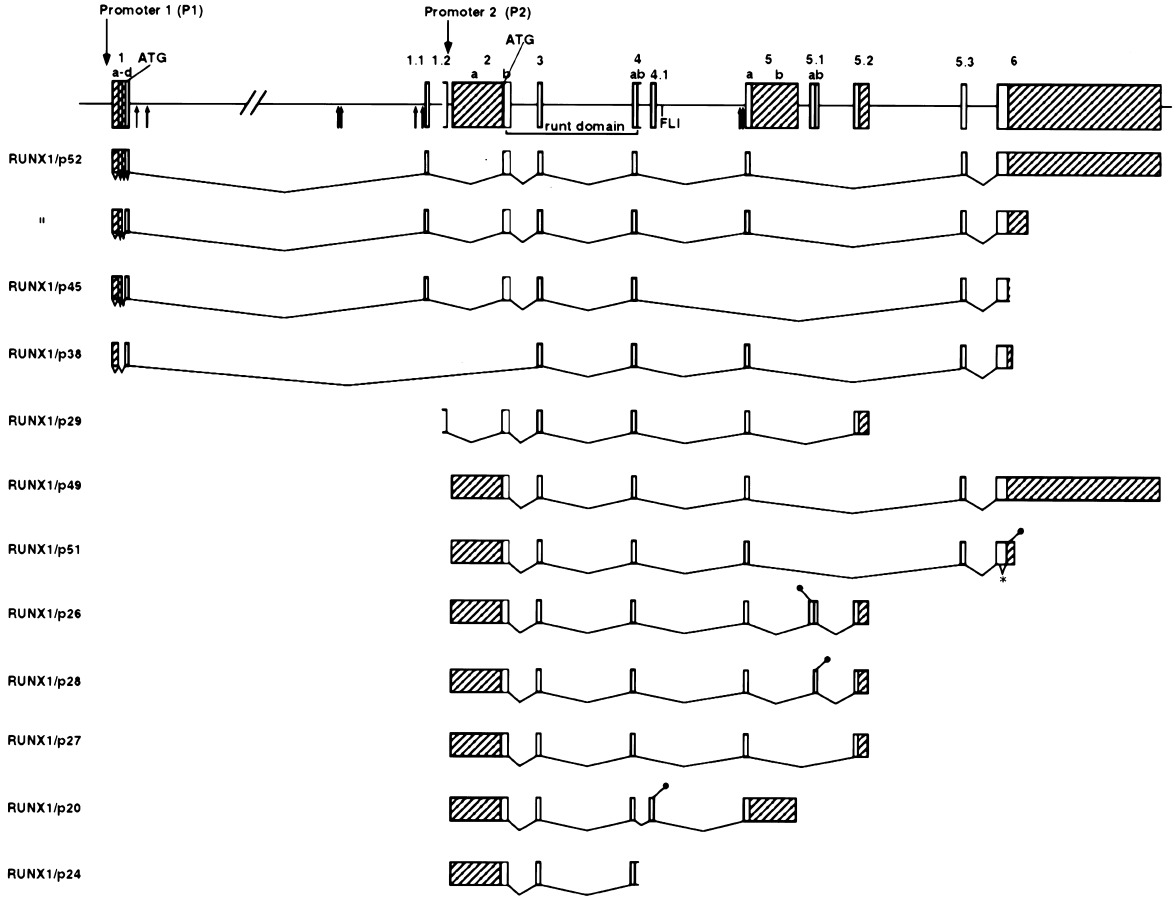


Fig. 1. Genomic organization of the human *RUNX1* gene (top line), and the structure of alternatively spliced mRNAs. Striped boxes represent UTRs. Shown cDNAs were isolated in our lab (Levanon et al., 1996), except: *RUNX1*/p52 (Miyoshi et al., 1995); *RUNX1*/p38 (Zhang et al., 1997); *RUNX1*/p27 (Miyoshi et al., 1991) and *RUNX1*/p24 (Sacchi et al., 1994). (●) marks in exon stop codon. (*) in the *RUNX1*/p51 cDNA denotes the deleted region. *RUNX1*/p29 and *RUNX1*/p24 are incomplete coding regions. The translocation breakpoints are marked by arrows beneath the gene map. The six t(12;21) breakpoints in intron 1 are according to Berger et al. (personal comunication) and Thandla et al. (Thandla et al., 1999). The t(8;21) and t(3;21) translocation breakpoints in intron 4.1 are according to (Shimizu et al., 1992; Hirai et al., 1999)

as in their exon composition. The latter give rise to a large repertoire of proteins, ranging in size between 20–52 kDa, that are generated through usage of alternatively spliced stop-codon-containing exons (marked by ● in Fig. 1). The genomic distance between the two 5′UTRs is 160 kb, in good correlation with previous estimates (Song et al., 1999; Thandla et al., 1999). The regions upstream of the 5′UTRs span the two *RUNX1* promoter regions P1 and P2 (Ghozi et al., 1996). Significantly, the two other human *RUNX* genes, *RUNX2* and *RUNX3*, have also two widely spaced transcriptional promoters (Stewart et al., 1997, our unpublished results). Transcription of *RUNX1* generates considerably long primary transcripts of either 260 kb (via P1) or 100 kb (via P2). The synthesis of these transcripts would take an estimated 2 h and 45 min respectively.

The exon/intron boundaries are listed in Table 1. While all splice junctions display the canonical GT/AG dinucleotides, the overall organization of the exons is quite complex (Fig. 1). Some exons (exon 4a and 5a) are either distinct (in most of the cDNA clones) or attached to a neighbouring exon (exon 4b and 5b-in clone *RUNX1*/p24 and *RUNX1*/p20, respectively). Of note, the 'b' parts of exons 4 and 5 does not appear as an independent exon, because the acceptor splice site are missing in both cases. An opposite situation occurs in exons 2 and 5.1. Exons 2b and 5.1b utilize a splice acceptor site at the end of part 'a' resulting in distinct exons 2b and 5.1b (cDNA *RUNX1*/p52 and *RUNX1*/p28, respectively). However, since 2b and 5.1b lack the donor splice site at their 5′ ends, their neighbouring exons (2a and 5.1a) never appear as distinct exons. An additional interesting feature was found in cDNA *RUNX1*/p51. Apart from a small deletion in the last exon, this clone is identical in its coding region to clone *RUNX1*/p49. A 99 bp deletion in the last exon of *RUNX1*/p49 (marked by * in Fig. 1) gives rise to clone *RUNX1*/p51 which differs from *RUNX1*/p49 in the C-terminal amino acids. This difference is biologically significant since clone *RUNX1*/p51 lacks the C-terminal motif VWRPY known to be involved in *RUNX1* mediated transcriptional repression (Fisher and Caudy, 1998; Levanon et al., 1998). The deleted region is not flanked by conventional donor/acceptor splice sites. Nevertheless, similar cDNA clones were isolated from a different cDNA library (Miyoshi et al., 1995). We therefore assume that clone *RUNX1*/p51 is generated by an unconventional splicing mechanism. It is known that besides the conventional pathway, splicing also occurs by a distinct type of spliceosome or even through spliceosome independent mechanisms (rev. in Abelson et al., 1998). Of note, the deleted 99 bp are extremely GC rich and capable of forming a stable hairpin structure bringing together the two ends of that region. This may facilitate an unconventional in-exon splicing event generating clone *RUNX1*/p51.

As mentioned above, expression of *RUNX1* is transcriptionally regulated by two promoters, giving rise to mRNAs bearing either UTR-1 or UTR-2 (Fig. 1). Recently we showed that these 5′UTRs act as translation regulators in vivo. UTR-1 mediates cap-dependent translation whereas the long structured UTR-2 contains an Internal Ribosomal Entry Site (IRES) and mediates cap-independent translation (Pozner et al., 2000). Hence, expression of *RUNX1* is regulated through usage of alternative promoters coupled with cap vs. IRES-mediated translation control. *RUNX1* expression is also regulated at the level of RNA splicing, as indicated by the large repertoire of mRNA species (Fig. 1). This multi level regulation of expression facilitates the generation of appropriate amounts of the relevant *RUNX1*

Table 1
Exon-intron boundaries of *RUNX1*[a]

| Exon no. | Size (bp) | Genomic location | Sequence |
|---|---|---|---|
| 1a | 176 | 40.798–40.973 | (prom)GAAAG...GCGTGgt |
| 1b | 88 | 40.974–41.061 | (1a)GTGAG...CAAAGgt |
| 1c | 123 | 41.061–41.183 | AgGTGCA...CGAAGgt |
| 1d | 117 | 41.183–41.299 | AgGTAAA...GAGAGgt |
| 1.1 | 39 | 197.181–197.219 | AgAATGC...CCACGgt |
| 1.2 | unknown 5′ end | 200.401–200.467 | (?)GCGAA...CCGGGgt |
| 2a | 1.597 | 201.451–203.047 | (prom)ATTCA...CGTAG(3b) |
| 2b | 254 | 203.048–203.301 | AgATGCC...TCAAGgt |
| 3 | 157 | 209.431–209.587 | AgGTGGT...AAGAGgt |
| 4a | 105 | 230.566–230.670 | AgGGAAA...TCGAAgt |
| 4b | unknown 3′ end | 230.671–230.810 | (5a)GTAAG...AAAGT(?) |
| 4.1 | 149 | 233.697–233.845 | AgACTCT...TTGAGgt |
| 5a | 192 | 255.543–255.734 | AgGACAT...GCAGGgt |
| 5b | 1.449 | 255.735–257.183 | (7a)GTAAG...AGAAG(polyA) |
| 5.1a | 52 | 260.173–260.224 | AgTTGTA...GATAG(8b) |
| 5.1b | 68 | 260.225–260.292 | AgACAAA...TGGAGgt |
| 5.2 | 417 | 268.448–268.864 | AgAGGAA...TCACT(polyA) |
| 5.3 | 162 | 290.682–290.843 | AgATACA...CTCAAgt |
| 6 | 4.797 | 297.534–302.331 | AgCGGCA...CAAGT(polyA) |

[a] Exon and intron sequences are shown in upper and lower case letters, respectively.

isoforms, at the proper time and in the correct cell type. Significantly, P1 derived transcripts give rise to UTR-1 bearing mRNAs with extended coding regions that include the TAD (Fig. 1). On the other hand, P2 derived transcripts bearing UTR-2 exhibit large variations in the coding regions due to alternative splicing (Fig. 1). Several of their coding regions are short and lack the TAD. The short protein isoforms bind to DNA with higher affinity than the full-length TAD containing proteins and are thought to act as dominant negative variants that out-compete the full length proteins for DNA binding (Tanaka et al., 1995; Ben Aziz-Aloya et al., 1998). Of note, the shorter transcription time of the P2 transcripts, relative to the P1, coupled with IRES mediated translation of P1 derived mRNAs, may influence the temporal and spatial production of the various *RUNX1* isoforms.

### 3.2. The genomic environment of the RUNX1 locus

A 1.87 Mb contig surrounding the genomic locus of the *RUNX1* gene was analyzed (Fig. 2). This genomic segment corresponds to the reverse-complement of range 21.0–22.9 Mb in the recently published DNA sequence (Hattori et al., 2000) and includes most of the Giemsa-dark chromosomal band 21q22.12. The emerging picture is of a sharp contrast between the genomic environments telomeric and centromeric to *RUNX1*, in the relative contents of G + C, CpG islands, repeats and genes. The region telomeric to *RUNX1* is a very large 'genomic graveyard' of imported pseudogenes (Fig. 2a). *RUNX1* appears to mark the transition between a very gene-poor, L isochore segment to an isochore H1 region that includes several genes.

Several genetic markers were located on it using e-PCR (rev. in Glusman and Lancet, 2000) and are displayed in Fig. 2b. An exhaustive search for neighboring genes by gene prediction (GenScan and fgenes) and homology searches (blastn/FASTA vs. GenBank and blastx/FASTX vs. GenPept) revealed the presence of at least five genes and five pseudogenes in addition to *RUNX1* (Fig. 2a). The apparently functional genes which are located centromeric to *RUNX1* include: *DSCR1*, two potassium channel regulatory subunit genes *KCNE1* and *KCNE2* of the IsK (minK) family, the sodium/myo-inositol cotransporter (*SLC5A3*) gene (Hattori et al., 2000) and a novel chloride channel

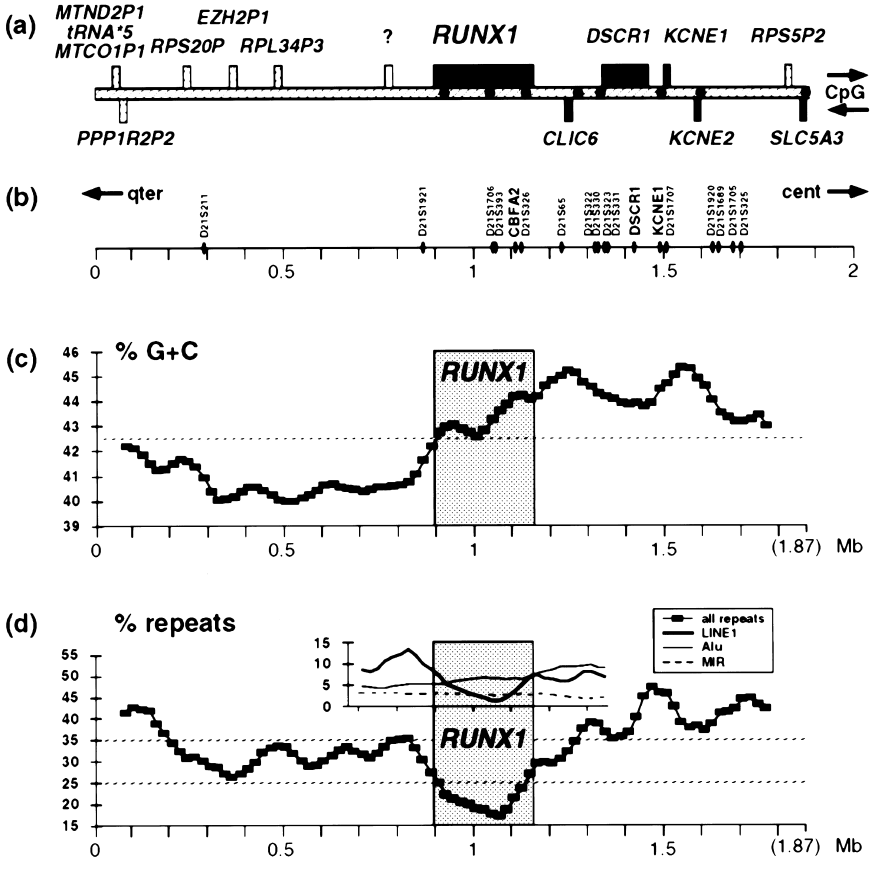


Fig. 2. Schematic map of *RUNX1* locus and its surroundings in megabases scale. (a) Gene map: boxes marked on top indicate genes in the same orientation as *RUNX1*; closed boxes indicate apparently functional genes. Large CpG islands are presented by dark circles in the midline. (b) Map of genetic markers in the region. Only D21S markers and those markers associated with genes are indicated. (c) G + C content throughout the region. The 42.5% line is shown for convenience. (d) Repeat content. The 25 and 35% lines are shown for convenience. The inset shows the partial repeat contents of LINE1, *Alu* and MIR in the immediate vicinity of *RUNX1*.

gene *CLIC*6 located 78.3 kb centromeric to *RUNX1* on the opposite strand. Each of these genes has an associated CpG island at its 5′ end (Fig. 2a). The region telomeric to *RUNX1* is highly populated with interspersed repeats and pseudogenes. The pseudogenes identified include: *PPP1R2P2*, apparently derived from the protein phosphatase 1 regulatory subunit 2 which maps to 3q29; *EZH2P1*, derived from the human homolog of the *Drosophila* enhancer of zeste gene *EZH2* which maps to 7q35; and 3 ribosomal protein pseudogenes, *RPS20P*, *RPL34P3* and *RPS5P2*. In addition, a highly diverged 1.8 kb segment of mitochondrial DNA sequence was observed (Fig. 2a), including two pseudogenes (*MTND2P1* and *MTCO1P1*) and five mitochondrial tRNA genes. For all the pseudogenes, the identity to the most similar paralog is higher at the nucleotide level than at the protein level, suggesting no selective pressure has acted on them. Some of them (e.g. *EZH2P1*) are clearly retrotransposed, processed pseudogenes. Notably, a trapped exon sequence with the GenBank accession HSZ98218 was found at positions 768,303–768,365 bp in the contig, 130 kb upstream of the first exon of *RUNX1*. This sequence displays stop codons in all three forward frames, though frame 3 could include an initial coding exon with the product MTEKLQTWAQ-. Exhaustive database searches found no significant similarities to other database sequences.

The analysis of the G + C content present in the 1.87 Mb contig (Fig. 2c) shows a sharp transition which coincides with the beginning of the *RUNX1* gene. Sequences telomeric to *RUNX1* correspond to an L isochore (40–42% G + C) and sequences centromeric from it correspond to an H1 isochore (43–45% G + C) (rev. in Glusman and Lancet, 2000). Some bias is also seen in the content of repetitive sequences (Fig. 2d), with sequences telomeric to *RUNX1* having a somewhat lower repeat content than sequences centromeric to it. A much stronger deviation in repeat content is seen along the gene itself (discussed in Section 3.5).

Human chromosome 21 has been reported to contain an unexpectedly low number of genes. Large gene-poor regions have been described (e.g. a 7 Mb region with only two known, and five predicted genes). These regions, which are concentrated in the centromeric half of the q arm of chromosome 21, tend to have low G + C content and are enriched in LINE1 repeats. The 1 Mb gene-less region described here has also G + C content lower than the average, but it is not enriched in LINE1 repeats. It is also the most telomeric of all such 'gene deserts' observed on chromosome 21, in sharp contrast to the gene-rich environment in which it is embedded. A comparison of the sequence-derived gene map of human chromosome 21 with the genetic map of mouse chromosome 16 (http://www.informatics.jax.org) shows that this 'gene desert' telomeric to *RUNX1* marks the boundary between a large region of conserved gene order (from *CBR1* to *MX1*) and a second region in which gene order is much less conserved (from *RUNX1* to *IL10RB* or *TIAM1*). In contrast to the observed gene order in human, mouse *Runx1* maps to the vicinity of *Tiam1*, about 5 cM centromeric to *Cbr*. Such a situation may be the result of a large inversion of up to 5 Mb between *CBR* and *TIAM1*, with later additional gene rearrangements. It is therefore tempting to speculate that the 1 Mb 'gene desert' described above was once a continuation of the very extensive, centromeric gene-poor sequence, with *RUNX1* being transcribed from centromere to telomere and much closer to the transition from the centromeric (gene-poor) to the telomeric (gene-rich) domain. The observed ancient mitochondrial segment marks the telomeric end of the 'gene desert', as well as the transition into 21q22.13, which has a higher G + C content and is richer in genes and *Alu* repeats. It is therefore an intriguing possibility that the insertion of the sequence derived from the mitochondrial genome may have triggered such an inversion in the mammalian lineage leading to primates, after the divergence from rodents.

### 3.3. Gene structure prediction

Fig. 3 displays the analysis of *RUNX1* DNA sequence using the GESTALT Workbench. The gene is located in an H1 isochore (coded blue in the %G + C graph). The only exception is the 3′ half of the long intron between P1 and P2 which has a lower G + C content than the rest of the gene. The *RUNX1* gene has 12 short coding exons. No other statistically significant contiguous open reading frames were observed, not even in the G + C rich regions of the CpG islands.

None of the gene prediction programs that we have applied (via GESTALT and RUMMAGE) modeled correctly the 160 kb long intron. Instead, several alternative gene structures were proposed, none of which yielded a product with significant similarities to sequences available in the databases. For the transcriptional unit starting at P2, the protein-coding part of the gene was modeled quite accurately by GenScan (Fig. 3c). However, three additional putative exons with relatively low scores were identified within two of the longest introns. Second-best results were obtained using XGrail. Regarding the P1 promoter, the experimentally identified exons got little support from any of the exon prediction programs used (only MZEF predicted exon 1.1). It is therefore apparent that these programs are less well suited for an accurate modeling of genes with characteristics like those of *RUNX1*, namely long introns and relatively short exons, some of which are non-coding. Under such circumstances, a combination of several programs may yield better results.

### 3.4. CpG islands

It is well documented that CpG dinucleotides are underrepresented in the human genome and appear in clusters (rev. in Glusman and Lancet, 2000). The 300 kb *RUNX1* sequence includes 22 CpG-enriched regions at least 200 bp long, comprising 3.7% of the entire gene sequence. Most of

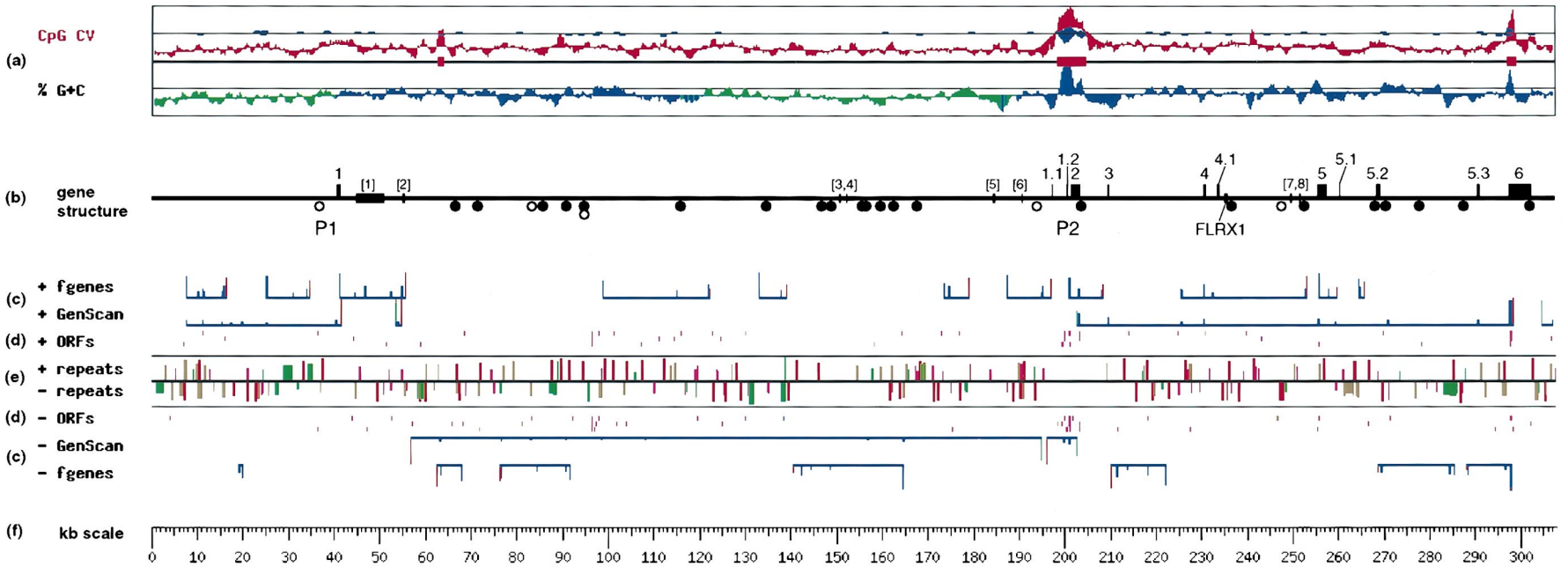


Fig. 3. A sequence map of *RUNX1* region generated by the GESTALT Workbench. (a) Compositional analyses. CpG contrast values (CV- the ratio between observed and expected frequencies, as CV = [CpG]/[C][G] where [C], [G] and [CpG] indicate frequencies) and %G + C are displayed as deviations from the regional average; large CpG islands are denoted by red boxes between the CpG CV and the %G + C graphs; Sp1 clusters are overlaid in blue; green %G + C stretches belong to the L isochore, blue stretches belong in the H1 isochore. (b) Gene structure: numbers on top denote the *RUNX1* exons; bracketed numbers indicate mapped breakpoints. Open circles represent chi sites (GCTGGTGG), and closed circles denote topoisomerase II consensus cleavage sites (RNYNNCNNGYNGKTNYNY). P1, P2 and FLRX1 represent the two promoters and the region of similarity to the *FLI1* gene. (c) Gene prediction results (fgenes and GenScan). Predicted exons are displayed in blue, with box height indicating exon quality (the scaling is arbitrary but consistent for each prediction program); complete gene structures are underlined in blue; promoters and polyA signals are indicated in green and red, respectively. (d) Location of open reading frames (ORFs), colors indicate statistical significance: brown and blue ORFs correspond to expectation value under $1 \times 10^{-3}$, respectively. (e) Repetitive sequences: *Alu*s are denoted in red, MIRs in purple, LINEs in green, other interspersed repeats in brown; box height indicates element age as percentage of identity with the subfamily consensus, from 50% (oldest) to 100% (youngest). (f) Kilobase scale. (c) to (e): Features on top of the middle line run from 5′ to 3′, and features under the middle line are in the reverse orientation.

Table 2
Location, length and repeat content of the *RUNX1* introns[a]

| Intron | Start | Length | % Repeat (all) | % Repeat (L1) | Similar length | % Lower (all) | % Lower (L1) |
|---|---|---|---|---|---|---|---|
| 1 | 41300 | 155881 | 20.42 | 3.11 | 11 | (18.18) | (9.09) |
| 1.1 | 197220 | 3181 | 0.00 | 0.00 | 3701 | 0.00 | 0.00 |
| 1.2 | 200468 | 983 | 0.00 | 0.00 | 4322 | 0.00 | 0.00 |
| 2 | 203302 | 6129 | 1.26 | 0.00 | 2333 | 1.50 | 0.00 |
| 3 | 209588 | 20978 | 17.13 | 0.00 | 564 | 6.74 | 0.00 |
| 4 | 230811 | 2886 | 2.91 | 0.00 | 3916 | 8.58 | 0.00 |
| 4.1 | 233846 | 21697 | 15.34 | 0.00 | 542 | 4.80 | 0.00 |
| 5 | 257184 | 2989 | 3.75 | 0.00 | 3862 | 8.80 | 0.00 |
| 5.1 | 260293 | 8155 | 46.03 | 0.00 | 1748 | 59.90 | 0.00 |
| 5.2 | 268865 | 21817 | 22.30 | 12.63 | 535 | 14.58 | 68.60 |
| 5.3 | 290844 | 6690 | 20.91 | 0.00 | 2143 | 20.58 | 0.00 |

[a] % lower indicates the fraction of database introns of similar length with lower repeat content (all repeats or L1 only). The values shown in brackets for intron 1 indicate statistical instability due to low sample size.

the 22 regions are short (<500 bp) and correspond to interspersed repeats of the *Alu* family. Three of them are large (790, 1060 and 3670 bp, Fig. 3a) and include clusters of Sp1 sites, known to function in the prevention of the re-methylation of CpG islands (rev. in Glusman and Lancet, 2000). The first two islands that span 3670 and 790 bp are located in the region of P2, at both ends of exon 2, with a stretch of 640 bp between them. An additional CpG island of 430 bp was detected 260 bp downstream of these two islands. Therefore, this region could be viewed as being a tripartite, 5.8 kb long CpG island. The third large (1060 bp) CpG island overlaps the beginning of the terminal exon 6. None of the other (shorter) potential CpG islands overlaps with any of the bona fide *RUNX1* exons. No CpG island was found at P1, the nearest one being within the first intron, 22 kb downstream of exon 1. The presence of a CpG island near the 3′ end of the gene is consistent with *RUNX1* being a tissue-specific gene (Gardiner-Garden and Frommer, 1987).

The largest *RUNX1* CpG island (3670 bp) is longer than any human CpG island in the database (CpGisle, version 4.0). The closest one is the 3340 bp island of the 28S ribosomal RNA gene, while the tissue specific expressed creatine kinase B gene possesses a much shorter one of 2432 bp. Our own global search for uninterrupted CpG-enriched segments in GenBank v.113 showed that among 5036 CpG islands which are longer than 700 bp, only 12 (0.24%) are longer than 3.67 kb, placing the *RUNX1* CpG island among the largest human CpG islands known.

### 3.5. Interspersed repeats

Visualization by GESTALT analysis of the *RUNX1* sequence suggests that overall the gene is relatively poor in repetitive sequences (Fig. 3e). Indeed, while the genomic regions surrounding the *RUNX1* gene contains on average 35% interspersed repeats (Fig. 2d), the gene itself contains only 19% repeats. This is mainly because the sequence is particularly poor in L1 repeats (Table 2, Fig. 2d), as seen by the fact that the overall repeat profile closely parallels that of L1 repeats (Fig. 2d, inset). The *Alu* repeats, on the other hand, are uniformly distributed throughout the gene.

A potential explanation for this distribution of repeats could be a dichotomy between intronic sequences (i.e. most of *RUNX1* sequence) and intergenic sequences: the latter may be more amenable to accepting retroposition events. To find out whether this is a feature of introns in general, or unique to the *RUNX1* introns, we performed a comparison to a collection of introns with similar lengths, derived from annotated GenBank sequences. The results suggest that all *RUNX1* introns, except intron 5.1, are specifically depleted of repeats (Table 2 and Fig. 4). This paucity is especially strong near P2 and in the region spanning 'runt domain' exons (introns 1.1–2).

From its isochore location, the sequence was expected to be enriched in L1 repeats. Therefore, the paucity of L1 repeats throughout the *RUNX1* transcriptional unit is not readily explainable. It may be speculated that a selection exists against further expansion of the already long introns by introduction of long insertions (up to 6 kb for L1). In addition, the open reading frames and control signals of L1 elements may disrupt the transcriptional unit. Noteworthy,

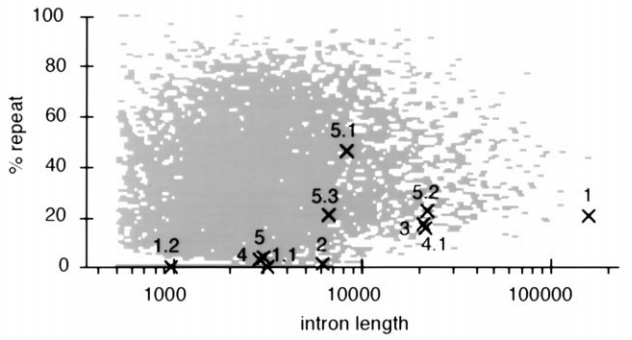


Fig. 4. Repeat content distribution in relation to intron length. Numbers (1–5.3) correspond to *RUNX1* introns, superimposed on information obtained from analyzing 8522 introns derived from GenBank sequences. Intron length is shown in logarithmic scale.

the few L1 inserts within the *RUNX1* introns are all oriented opposite to the transcription of the gene.

### 3.6. Breakpoints of leukemia associated chromosomal translocations

The leukemia associated translocations interrupt the *RUNX1* gene at several positions. The breakpoints of the t(8;21) and t(3;21) translocations were mapped to introns 4, 4.1, 5 and 5.1 (Figs. 1 and 3) (Nucifora and Rowley, 1995). In these cases, the resulting chimeric genes are regulated by the two *RUNX1* promoters since they include the 5′ regulatory regions of *RUNX1* fused to the partner gene. The t(12;21) translocation breakpoints occur in the large intron between the two promoters (Figs. 1 and 3). In this translocation the *TEL* gene from chromosome 12 is fused to the *RUNX1* locus downstream to P1. Transcription of the fused gene is therefore regulated by the *TEL* promoter and by the *RUNX1* P2 promoter (Thandla et al., 1999 and refernences therein). As seen in Figs. 1 and 3, the reported breakpoints in the t(12;21) translocation are clustered in three regions within the long intron #1. Also shown are the reported intronic breakpoint for the t(8;21) translocation (Shimizu et al., 1992) and in its vicinity the breakpoint for the t(3;21) translocation (Hirai et al., 1999).

Recent reports showed that both chi sites and topoisomerase II binding and cleavage sites are found at or near translocation breakpoints (Hirai et al., 1999 and refs. therein). Such sites typically did not perfectly match with the published consensi (GCTGGTGG and RNYNNCNNGYNGKTNYNY, respectively), but rather showed up to two mismatches. We now performed a search for chi sites and topoisomerase II cleavage sites along the 300 kb sequence of *RUNX1*. Five and 22 perfectly matching sites were observed respectively, in good agreement with the expectation based on the G + C content (5.1 and 27.6, respectively). No clustering of chi sites is apparent (Fig. 3). Interestingly, all the observed topoisomerase II cleavage sites are located within the transcription unit (none in the 40 kb upstream of exon 1), but they are not randomly distributed: there is a cluster of seven sites in a 25 kb segment within long intron #1. This segment includes breakpoints three and four (Fig. 3). This clustering is intriguing since this segment has the lowest G + C content within *RUNX1* transcription unit (i.e. less sites are expected). A search for topoisomerase II cleavage sites, allowing up to two mismatches from the consensus, shows a large number of such sequences at the large CpG island complex, even after normalization by G + C content.

### 3.7. RUNX1/FLI1 homology

A region of 555 bp (hereafter called FLRX1) located within intron 4.1 (at position 235,178–235,732 bp), where the t(8;21) translocation breakpoints occur (Nucifora and Rowley, 1995), was found to share a high degree of identity with an intronic region of the *FLI1* gene (between exon 3 and 4, at position 10239–10786 bp in GenBank entry HSY17293; Fig. 5a). *FLI1* is located on chromosome 11 and participates in the t(11;22) translocation generating the EWS/FLI1 fusion transcript associated with Ewing's tumors (Delattre et al., 1992). Interestingly, the *FLI1* sequence which shares similarity with *RUNX1* harbors one of the t(11;22) translocation breakpoints (GenBank entry HAJ9349), and is located at the edge of a ~40 kb region which accomodates all the other EWS/FLI translocation breakpoints (Zucman-Rossi et al., 1998). Most of FLRX1 (351 bp) does not correspond to any family of interspersed repeats. The remaining 204 bp belong to an *Alu*Sx repeat. Of note, while in *RUNX1* this *Alu* element is truncated, the complete element of 303 bp is present in *FLI1*, flanked by short direct repeats that are generated during the retroposi-

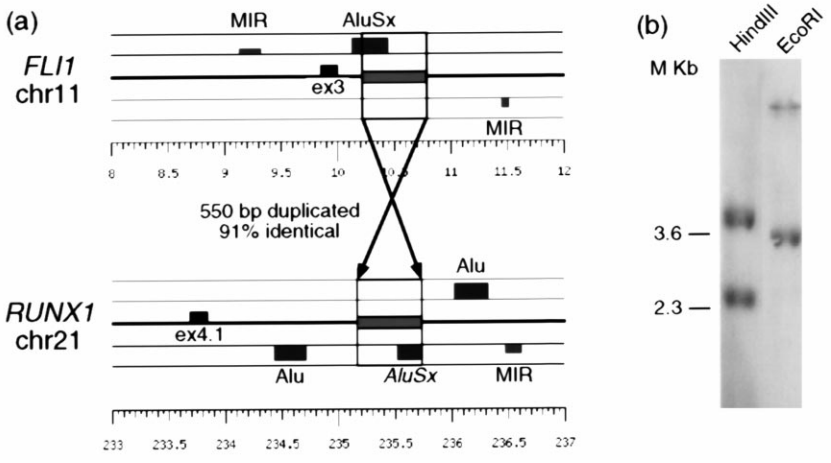


Fig. 5. (a) GESTALT view of the *FLI1*/*RUNX1* common region. For simplicity, the scale is presented in kb only, the scheme includes only repeats and user annotation (for details see legend to Fig. 2). The common region is striped. Crossing arrows indicate the extent and orientation of the copied region. (b) Southern blot analysis. Twenty micrograms of human DNA was digested with either HindIII or EcoRI. Blot was hybridized under stringent conditions to random primed probe prepared from the 350 bp intronic *RUNX1* fragment with high sequence similarity to *FLI1*.

tion event of *Alu* repeat. We therefore concluded that FLRX1 has originated in *FLI1* and was 'imported' into *RUNX1*. From the level of divergence (91% identity), this transposition event is estimated to have taken place 25–35 Myr ago.

The significance of this finding is highlighted by the fact that *RUNX1* and *FLI1* share several common features. Similar to *RUNX1, FLI1* (that belongs to the Ets gene family) encodes a hematopoietic transcription factor. Expression of *FLI1* also involves two distinct promoters and shows elaborate alternative splicing (Dhulipala et al., 1998). Finally, *FLI1* is also involved in oncogenic translocations. This intriguing sequence similarity between two genomic regions that are involved in malignant transformation-associated translocations prompted us to analyze whether additional copies of a similar sequence exist in the genome. For this purpose, we cloned the chromosome 21 specific sequence (Fig. 5a) and used it as a probe on a genomic Southern blot. Two hybridizing bands were detected in either HindIII or EcoRI digested human DNA (Fig. 5b). Based on their indicated sizes, the upper band in the HindIII digest and the lower band in the EcoRI digest, represent genuine chromosome 21 sequences (Fig. 5b). This indicates that there are only two copies of the above sequence in the genome. However, whether in *RUNX1* and *FLI1* these sequences contribute to the high incidence of oncogenesis associated translocations remains an open question.

## Acknowledgements

This work was supported by grants from the Commission of the European Community's Biomedicine and Health research program BIOMED II No. PL963039 and the Shapell Family Biomedical Research Foundation at the Weizmann Institute.

## References

Abelson, J., Trotta, C.R., Li, H., 1998. tRNA splicing. J. Biol. Chem. 273, 12685–12688.

Ben Aziz-Aloya, R., Levanon, D., Karn, H., Kidron, D., Goldenberg, D., Lotem, J., Polak-Chaklon, S., Groner, Y., 1998. Expression of AML1-d, a short human AML1 isoform, in embryonic stem cells suppresses in vivo tumor growth and differentiation. Cell Death Differ. Cell Death Differ 5, 765–773.

Delattre, O., Zucman, J., Plougastel, B., Desmaze, C., Melot, T., Peter, M., Kovar, H., Joubert, I., de Jong, P., Rouleau, G., 1992. Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours. Nature 359, 162–165.

Dhulipala, P.D., Lee, L., Rao, V.N., Reddy, E.S., 1998. Fli-1b is generated by usage of differential splicing and alternative promoter. Oncogene 17, 1149–1157.

Downing, J.R., 1999. The AML1-ETO chimaeric transcription factor in acute myeloid leukemia: biology and clinical significance. Br. J. Haematol. 106, 296–308.

Fisher, A.L., Caudy, M., 1998. Groucho proteins: transcriptional corepressors for specific subsets of DNA-binding transcription factors in vertebrates and invertebrates. Genes Dev. 12, 1931–1940.

Gardiner-Garden, M., Frommer, M., 1987. CpG Islands in Vertebrate Genomes. J. Mol. Biol. 196, 261–282.

Ghozi, M.C., Bernstein, Y., Negreanu, V., Levanon, D., Groner, Y., 1996. Expression of the human acute myeloid leukemia gene AML1 is regulated by two promoter regions. Proc. Natl. Acad. Sci. USA 93, 1935–1940.

Glusman, G., Lancet, D., 2000. GESTALT: a workbench for automatic integration and visualization of large-scale genomic sequence analyses. Bioinformatics 16, 482–483.

Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H., Yada, T., Park, H.S., Toyoda, A., Ishii, K., Totoki, Y., Groner, Y., Choi, D.K., Soeda, E., Ohki, M., Takagi, T., Sakaki, Y., Taudien, S., Blechschmidt, K., Polley, A., Menzel, U., Delabar, J., Kumpf, K., Lehmann, R., Patterson, D., Reichwald, K., Rump, A., Schillhabel, M., Schudy, A., 2000. The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. Nature 405, 311–319.

Hirai, H., Ogawa, S., Kurokawa, M., Yazaki, Y., Mitani, K., 1999. Molecular characterization of the genomic breakpoints in a case of t(3;21)(q26;q22). Genes Chromosomes Cancer 26, 92–96.

Ito, Y., 1999. Molecular basis of tissue-specific gene expression mediated by the runt domain transcription factor PEBP2/CBF. Genes Cells 4, 685–696.

Ito, Y., Bae, S.C., 1997. The Runt domain transcription factor, PEBP2/CBF, and its involvement in human leukemia. In: Yaniv, M., Ghysdael, J. (Eds.), Oncogenes as Transcriptional Regulators. Birkhauser Verlag, Basel, Switzerland, pp. 107–132.

Levanon, D., Bernstein, Y., Negreanu, V., Ghozi, M.C., Bar-Am, I., Aloya, R., Goldenberg, D., Lotem, J., Groner, Y., 1996. A large variety of alternatively spliced and differentially expressed mRNAs are encoded by the human acute myeloid leukemia gene AML1. DNA Cell Biol. 15, 175–185.

Levanon, D., Goldstein, R.E., Bernstein, Y., Tang, H., Goldenberg, D., Stifani, S., Paroush, Z., Groner, Y., 1998. Transcriptional repression by AML1 and LEF-1 is mediated by the TLE/Groucho corepressors. Proc. Natl. Acad. Sci. USA 2995, 11590–11595.

Levanon, D., Negreanu, V., Bernstein, Y., Bar-Am, I., Avivi, L., Groner, Y., 1994. AML1 AML2, and AML3, the human members of the runt domain gene-family: cDNA structure, expression, and chromosomal localization. Genomics 23, 425–432.

Look, A.T., 1997. Oncogenic transcription factors in the human acute leukemias. Science 278, 1059–1064.

Meyers, S., Lenny, N., Sun, W.H., Hiebert, S.W., 1996. AML-2 is a potential target for transcriptional regulation by the t(8;21) and t(12;21) fusion proteins in acute leukemia. Oncogene 13, 303–312.

Miyoshi, H., Ohira, M., Shimizu, K., Hirai, H., Imai, T., Yokoyama, K., Soeda, E., Ohki, M., 1995. Alternative splicing and genomic structure of the *AML1* gene involved in acute myeloid leukemia. Nucleic Acids Res. 23, 2762–2769.

Miyoshi, H., Shimizu, K., Kozu, T., Maseki, N., Kaneko, Y., Ohki, M., 1991. t(8;21) breakpoints on chromosome 21 in acute myeloid leukemia are clustered within a limited region of a single gene. *AML1*. Proc. Natl. Acad. Sci. USA 88, 10431–10434.

Nucifora, G., Rowley, J.D., 1995. *AML1* and the 8;21 and 3;21 translocations in acute and chronic myeloid leukemia. Blood 86, 1–14.

Pozner, A., Goldenberg, D., Negreanu, V., Le, S.-Y., Elroy-Stein, O., Levanon, D., Groner, Y., 2000. Transcription-coupled translation control of AML1/*RUNX1* is mediated by cap- and internal ribosome entry site-dependent mechanisms. Mol. Cell. Biol. 20, 2297–2307.

Sacchi, N., Nisson, P.E., Watkins, P.C., Faustinella, F., Wijsman, J., Hagemeijer, A., 1994. AML1 fusion transcripts in t(3;21) positive leukemia: evidence of molecular heterogeneity and usage of splicing sites frequently involved in the generation of normal AML1 transcripts. Genes, Chromosomes Cancer 11, 226–236.

Satake, M., Nomura, S., Yamaguchi-Iwai, Y., Takahama, Y., Hashimoto, Y., Niki, M., Kitamura, Y., Ito, Y., 1995. Expression of the runt domain-encoding PEBP2α genes in T cells during thymic development. Mol. Cell Biol. 15, 1662–1670.

Shimizu, K., Miyoshi, H., Kozu, T., Nagata, J., Enomoto, K., Maseki, N., Kaneko, Y., Ohki, M., 1992. Consistent disruption of the AML1 gene occurs within a single intron in the t(8;21) chromosomal translocation. Cancer Res. 52, 6945–6948.

Song, W.J., Sullivan, M.G., Legare, R.D., Hutchings, S., Tan, X., Kufrin, D., Ratajczak, J., Resende, I.C., Haworth, C., Hock, R., Loh, M., Felix, C., Roy, D.C., Busque, L., Kurnit, D., Willman, C., Gewirtz, A.M., Speck, N.A., Bushweller, J.H., Li, F.P., Gardiner, K., Poncz, M., Maris, J.M., Gilliland, D.G., 1999. Haploinsufficiency of *CBFA2* causes familial thrombocytopenia with propensity to develop acute myelogenous leukamia. Nat. Genet. 23, 166–175.

Speck, N.A., Stacy, T., Wang, Q., North, T., Gu, T.L., Miller, J., Binder, M., Marin-Padilla, M., 1999. Core-binding factor: a central player in hematopoiesis and leukemia. Cancer Res 59, 1789s–1793s.

Stewart, M., Terry, A., Hu, M., O'Hara, M., Blyth, K., Baxter, E., Cameron, E., Onions, D.E., Neil, J.C., 1997. Proviral insertions induce the expression of bone-specific isoforms of PEBP2alphaA (CBFA1): evidence for a new myc collaborating oncogene. Proc. Natl. Acad. Sci. USA 94, 8646–8651.

Tanaka, T., Tanaka, K., Ogawa, S., Korokawa, M., Mitani, K., Nishida, J., Shibata, Y., Yazaki, Y., Hirai, H., 1995. An acute myeloid leukemia gene AML1, regulates hemopoietic myeloid cell differentiation and transcriptional activation antagonistically by two alternative spliced forms. EMBO J. 14, 341–350.

Thandla, S.P., Ploski, J.E., Raza-Egilmez, S.Z., Chhalliyil, P.P., Block, A.W., de Jong, P.J., Aplan, P.D., 1999. ETV6-AML1 translocation breakpoints cluster near a purine/pyrimidine repeat region in the ETV6 gene. Blood 93, 293–299.

Warren, A.J., Bravo, J., Williams, R.L., Rabbitts, T.H., 2000. Structural basis for the heterodimeric interaction between the acute leukaemia-associated transcription factors AML1 and CBFbeta. EMBO J. 19, 3004–3015.

Zhang, Y.M., Bae, S.C., Huang, G., Fu, Y.X., Lu, J., Ahn, M.Y., Kanno, Y., Kanno, T., Ito, Y., 1997. A novel transcript encoding an N-terminally truncated AML1/PEBP2αB protein interferes with transactivation and blocks granulocytic differentiation of 32Dc13 myeloid cell. EMBO J. 14, 341–350.

Zucman-Rossi, J., Legoix, P., Victor, J.M., Lopez, B., Thomas, G., 1998. Chromosome translocation based on illegitimate recombination in human tumors. Proc. Natl. Acad. Sci. USA 95, 11786–11791.