Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 3623–3627, June 1983
Biochemistry

Nucleotide sequence of Abelson murine leukemia virus genome: Structural similarity of its transforming gene product to other *onc* gene products with tyrosine-specific kinase activity

(open reading frame / protein homology)

E. PREMKUMAR REDDY, M. J. SMITH, AND A. SRINIVASAN

Laboratory of Cellular and Molecular Biology, National Cancer Institute, Building 37, Room 1A07, Bethesda, Maryland 20205

Communicated by Michael Potter, March 14, 1983

ABSTRACT **The nucleotide sequence of the proviral genome of Abelson murine leukemia virus (A-MuLV), an acute transforming virus of murine origin, has been determined. Like other transforming viruses, A-MuLV contains sequences derived from its helper virus, Moloney murine leukemia virus (M-MuLV), and a cell-derived protooncogene (*abl*) insertion sequence. By comparison of the A-MuLV sequence with that of M-MuLV, it was possible to precisely localize and define sequences contributed by the host cellular DNA. From the nucleotide sequence, we have predicted the amino acid sequence of p120$^{gag-abl}$, the product of the A-MuLV *gag–abl* hybrid gene. The amino acid sequence of the putative *abl* gene, when compared with the sequences of other tyrosine-specific protein kinases (*src*, *fes*, *fps*, and *yes*), revealed significant homologies, indicating that all these functionally related transforming genes are derived from divergent members of the same protooncogene family. In addition to the *gag–abl* sequence, the proviral genome was found to contain an additional open reading frame that could code for an 18,000-dalton protein, whose role is at present undetermined.**

Abelson murine leukemia virus (A-MuLV) is a replication-defective transforming retrovirus that was isolated after inoculation of Moloney murine leukemia virus (M-MuLV) in prednisolone-treated BALB/c mice (1). A-MuLV induces B-cell lymphomas *in vivo* (2, 3) and is able to transform both lymphoid and fibroblastic cells *in vitro* (for review, see ref. 4). Several lines of evidence indicate that A-MuLV arose by recombination of the nondefective helper virus (M-MuLV) and cellular sequences present within the normal mouse genome (5–7). The latter sequences, termed *abl*, appear to code for the transforming properties of the virus (4, 7).

A-MuLV transforming functions are thought to be mediated by a 120,000-dalton polyprotein, p120. This protein is a hybrid molecule containing M-MuLV *gag* gene structural proteins as well as a MuLV-unrelated component coded by *abl* sequences (8). p120 coded by the A-MuLV genome has been shown to possess closely associated kinase activity with specificity for tyrosine phosphorylation (9). In an effort to understand the mechanism of action of the v-*abl* gene, and to understand its relationship with other transforming genes known to code for tyrosine-specific kinase activity, we have undertaken primary DNA sequence analysis of the molecularly cloned integrated proviral genome (6, 7). Contrary to previous evidence from molecular hybridization studies, p120 was found to have significant homology with the transforming gene product of retroviruses [Rous sarcoma virus (RSV), Y73, feline sarcoma virus (FeSV), and Fujinami sarcoma virus (FSV)] coding for tyrosine-specific kinase activity.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

MATERIALS AND METHODS

Molecular Cloning of A-MuLV DNA. The isolation of a molecular clone of integrated A-MuLV DNA from mink cells nonproductively transformed by A-MuLV in λgt*WES*·λB (6) and its subsequent subcloning in pBR322 (7) have been described. The A-MuLV DNA insert from plasmid DNA was purified by agarose gel electrophoresis and chromatography on DEAE-cellulose (DE-52, Whatman) and used in all subsequent analysis.

DNA Sequence Analysis. The nucleotide sequence was determined by the procedure described by Maxam and Gilbert (10). DNA fragments were obtained by using various restriction endonucleases and were labeled either at their 5′ end by using [γ-^{32}P]ATP (Amersham, 3,000 Ci/mmol; 1 Ci = 3.7×10^{10} Bq) and polynucleotide kinase (P-L Biochemicals) (10) or at their 3′ end by using cordycepin 5′-[α-^{32}P]triphosphate (Amersham, 3,000 Ci/mmol) and terminal deoxynucleotidyltransferase (P-L Biochemicals) according to the method of Roychoudhury and Wu (11). End-labeled DNA fragments were digested with appropriate restriction endonucleases (New England BioLabs), isolated by agarose or polyacrylamide gel electrophoresis, and used for sequence analysis.

RESULTS

Nucleotide Sequence of A-MuLV. The primary nucleotide sequence of the integrated A-MuLV genome was determined according to the partial chemical degradation method of Maxam and Gilbert (10). The sequence for both strands was determined for most of the genome, and known restriction cleavage sites were confirmed by sequence analysis. The complete sequence of A-MuLV along with its flanking mink cellular sequences are presented in Fig. 1. Fig. 2 provides a summary of some of the salient features of the viral genome.

Identification of M-MuLV and v-*abl* Coding Sequences Within the A-MuLV Genome. Because the complete nucleotide sequence of M-MuLV, the natural helper virus of A-MuLV, is known (12), we compared the sequence presented in Fig. 1 with that of M-MuLV. Such a comparison revealed that the sequence presented in Fig. 1 contained 79- and 156-base-pair (bp) flanking mink cellular sequences at its 5′ and 3′ ends, respectively. These mink cellular sequences immediately flanking the A-MuLV LTRs were found to contain a four-nucleotide direct repeat sequence, T-G-G-G, confirming previous findings of duplication of a short stretch of sequences at the site of retrovirus integration (13, 14).

The comparison also revealed the occurrence of a sequence homology of 1,776 bp at the 5′ end and 793 bp at the 3′ end.

Abbreviations: A-MuLV and M-MuLV, Abelson and Moloney murine leukemia virus; RSV, Rous sarcoma virus; MuSV, murine sarcoma virus; bp, base pair(s); LTR, long terminal repeat.

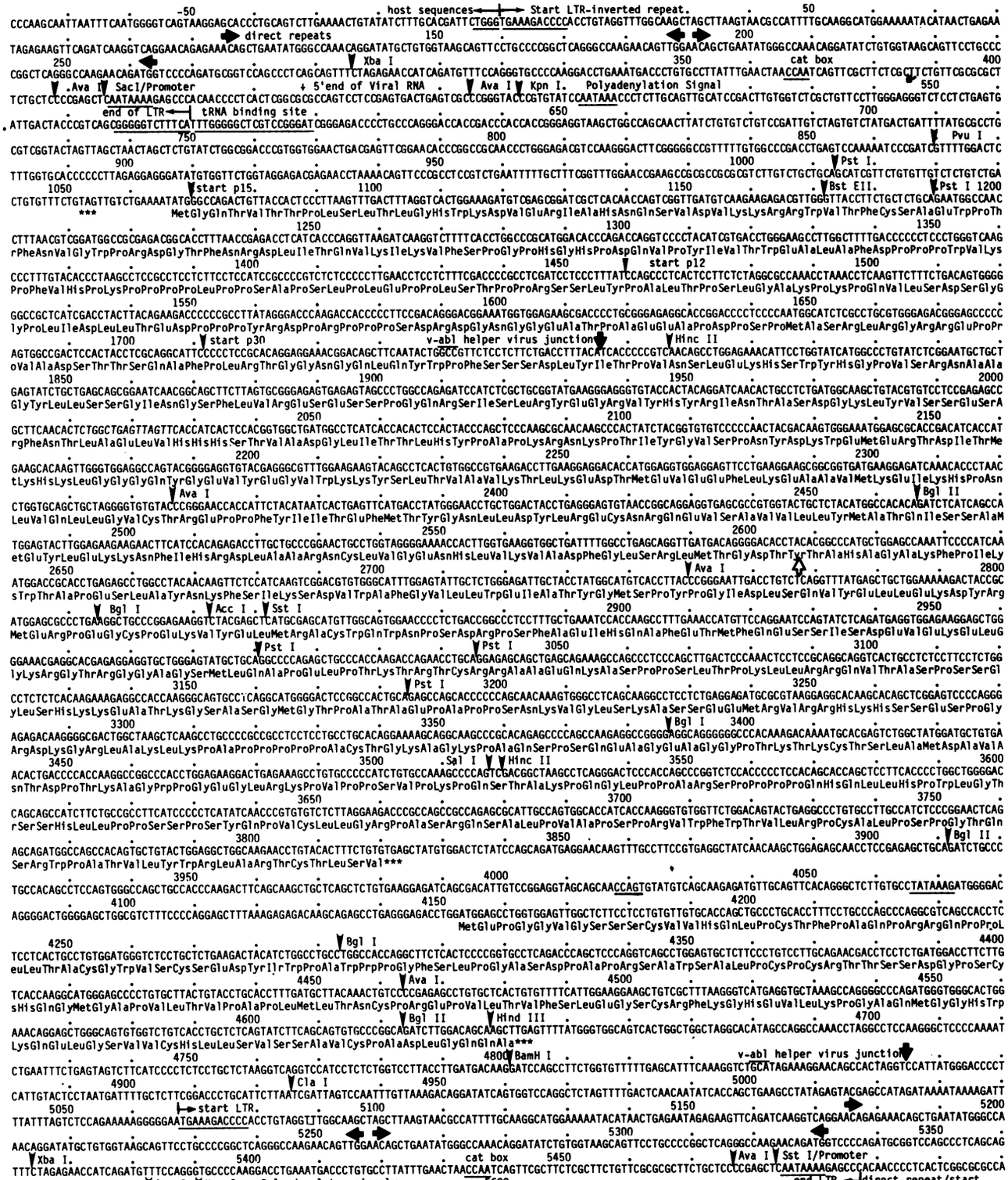

FIG. 1. Nucleotide sequence of the proviral A-MuLV genome. The sequence proceeding in the 5′-to-3′ direction has the same polarity as A-MuLV genomic RNA. Dots mark every 10th nucleotide. The amino acid sequence deduced from the open reading frame is given below the nucleotide sequence. The major structural features of the genome are indicated. LTR, long terminal repeat.

The region of homology at the 5′ end included the noncoding sequence and the amino-terminal region of the *gag* gene. The noncoding sequence at the 5′ terminus included the LTR, the primer tRNA binding site, and a stretch of sequences that are present before the start of the *gag* sequences. This region of 1,776 bp exhibited 17 differences between the two genomes. Eight of these changes occurred in the LTR region, three in the 5′ noncoding sequences, and six within the *gag* gene. The open reading frame, starting at position 1,067, contained the entire sequence of p15 and p12 and the first 21 codons of p30. Beyond position 1,776, no sequence homology was observed between the two viral genes, thus localizing the point of recombination

*Correction - Proc. Natl. Sci. U.S.A. 80 (1983) p. 7372

Correction. In the article "Nucleotide sequence of Abelson murine leukemia virus genome: Structural similarity of its transforming gene product to other *onc* gene products with tyrosine-specific kinase activity" by E. Premkumar Reddy, M. J. Smith, and A. Srinivasan, which appeared in number 12, June 1983, of *Proc. Natl. Acad. Sci. USA* (80, 3623–3627), the authors request the following correction. The nucleotide sequence in Fig. 1 was found to be missing a single cytosine residue at position 3,534. Correction of this sequence is presented below. This correction alters the reading frame toward the carboxyl terminus of the open reading frame presented in the new figure. Since this correction occurs in the region of the v-*abl* genome that is not essential for fibroblast transforming activity and has no counterpart in pp60 *src*, none of the major conclusions derived from the sequence are affected.

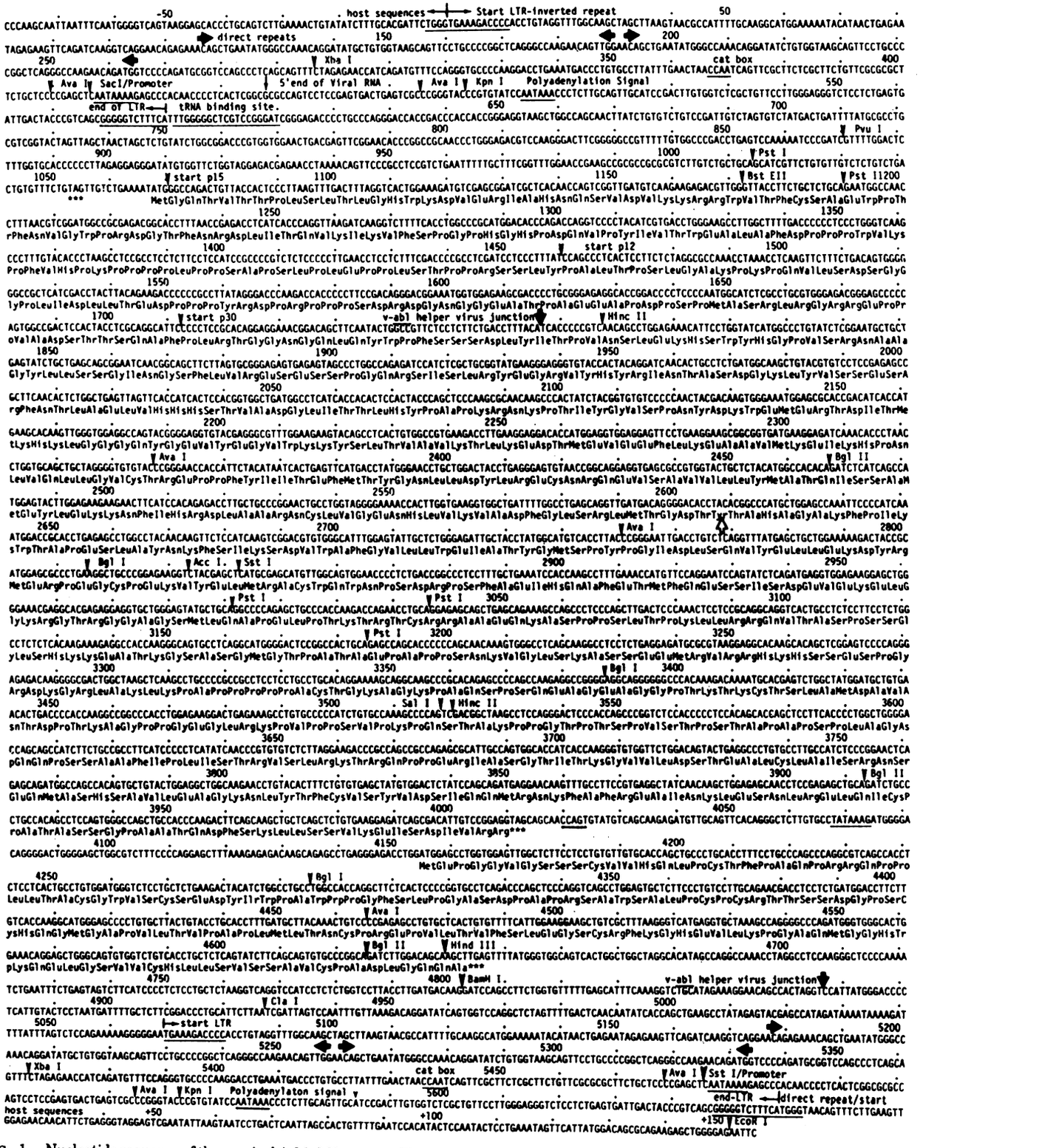

FIG. 1. Nucleotide sequence of the proviral A-MuLV genome. The sequence proceeding in the 5′-to-3′ direction has the same polarity as A-MuLV genomic RNA. Dots mark every 10th nucleotide. The amino acid sequence deduced from the open reading frame is given below the nucleotide sequence. The major structural features of the genome are indicated. LTR, long terminal repeat.

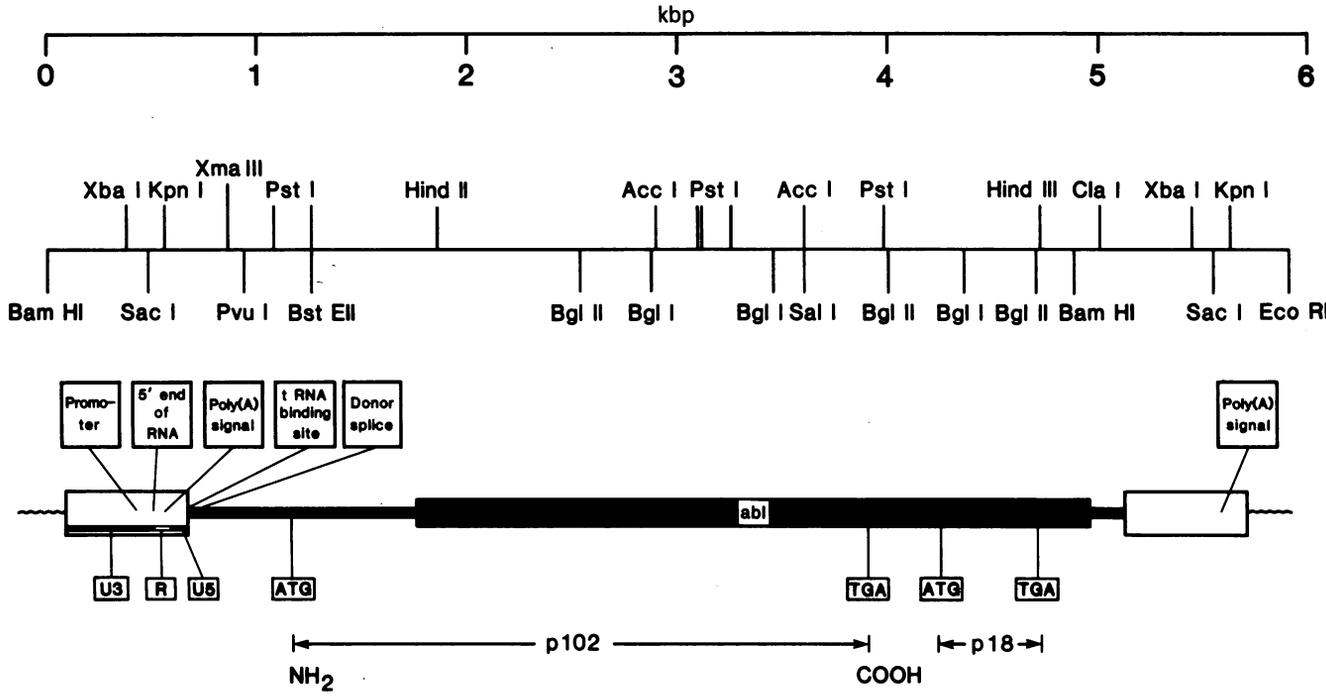

FIG. 2. Summary of the major structural features of the A-MuLV genome. Important features of A-MuLV genome, including the open reading frames, possible signals for promoter, poly(A) addition, and donor and acceptor splice signals are illustrated. kbp, Kilobase pairs.

at the 5′ end. It is interesting to note that Goff and Baltimore (15) reported the occurrence of 5-bp homology between the helper viral and c-*abl* sequences at the point of recombination, leading these authors to speculate on the possibility of homologous recombination between the two genes. Such homologies were also found to occur in Moloney murine sarcoma virus (MuSV), simian sarcoma virus, and myelocytomatosis (MC29) viral genes (16–18). Beyond the point of recombination, the open reading frame extended for an additional stretch of 2,045 bp, terminating with a TGA codon. The open reading frame shown in Fig. 1 could code for a polypeptide of 918 amino acids with an approximate molecular mass of 102,000 daltons. This is in reasonable agreement to the estimated size of 110,000–120,000 daltons for the *gag–abl* hybrid protein synthesized by A-MuLV-infected NIH/3T3 nonproducer cells. This protein consists of 240 amino acids derived from the amino terminus of the *gag* region followed by 678 amino acids that are specific to the v-*abl* region. The coding sequences terminated within the cell-derived *abl* sequences 800 bases upstream from the v-*abl* helper viral junction at the 3′ end. Examination of this additional stretch of 800 bp revealed the presence of a second open reading frame starting with an ATG at position 4,153–4,155 and terminating at position 4,642–4,644 with a TGA codon. This stretch of 492 bases could code for a protein of 18,000 daltons. It is interesting to note that a promoter-like sequence T-A-T-A-A-A occurs at position 4,066–4,071 upstream to this open reading frame. A sequence resembling C-C-A-A-T box is also seen at position 4,018–4,022.

The point of recombination between the M-MuLV and c-*abl* sequences at the 3′ end occurred at the same point within the viral genome as it did with c-*mos* during the generation of the Moloney MuSV genome (16, 19, 20). These findings suggest that there may exist preferential sites for recombination within the helper viral genome, and these sites could have played a crucial role in their evolution.

Amino Acid Sequence Homology with Other Viral Oncogenes Coding for Tyrosine-Specific Kinase Activity. The translational product of the A-MuLV genome, P120, has been shown to be associated with tyrosine kinase activity (9). Of about 15 described v-*onc* genes, at least 6 appear to encode enzymes with analogous function (*src*, *yes*, *fes*, *fps*, *ros*, and *abl*) (for review see ref. 21). It was, therefore, interesting to examine the structural relationships among these *onc* proteins because the sequence for five of them is available (22–25). Such a comparative analysis with $pp60^{src}$ is presented in Fig. 3. The results of this analysis demonstrated that all these five proteins have extensive sequence homology. Thus, v-*abl* protein shared 176, 166, 140, and 138 amino acids with *yes*-, *src*-, *fes*-, and *fps*-encoded transforming proteins (22–25). The homology is more pronounced with regions that are implicated with the active site for tyrosine phosphorylation (22–25). Less conserved regions of homology are observed between v-*abl* and v-*mos* encoded gene products (16, 20); one of the v-*mos* products also has been shown to be a protein kinase (27).

DISCUSSION

Nucleotide sequence analysis of the A-MuLV proviral genome has revealed several important features of its molecular organization. Examination of the sequence data presented here reveals a large open reading frame on the viral RNA strand which could code for a protein of 918 amino acids with a molecular mass of 102,000 daltons.

Like many of the transforming viruses, A-MuLV appears to synthesize its transforming protein by means of a *gag–abl* polyprotein, the amino-terminal region of which is composed of helper virus *gag* gene products. In the case of A-MuLV, the *gag–abl* hybrid protein contains the entire sequence of p15 and p12 and the first 21 amino acids of p30 and 678 amino acids derived from v-*abl*. Thus, the transforming protein utilized the helper viral sequences for the initiation of its synthesis. Earlier studies with M-MuLV had demonstrated that the product *gag* gene is formed by processing of a larger precursor polypeptide (28), which is approximately 6,000 daltons larger than the 65,000-dalton *gag* polyprotein. Sequence analysis of the cloned viral DNAs of M-MuLV and Moloney MuSV revealed the presence

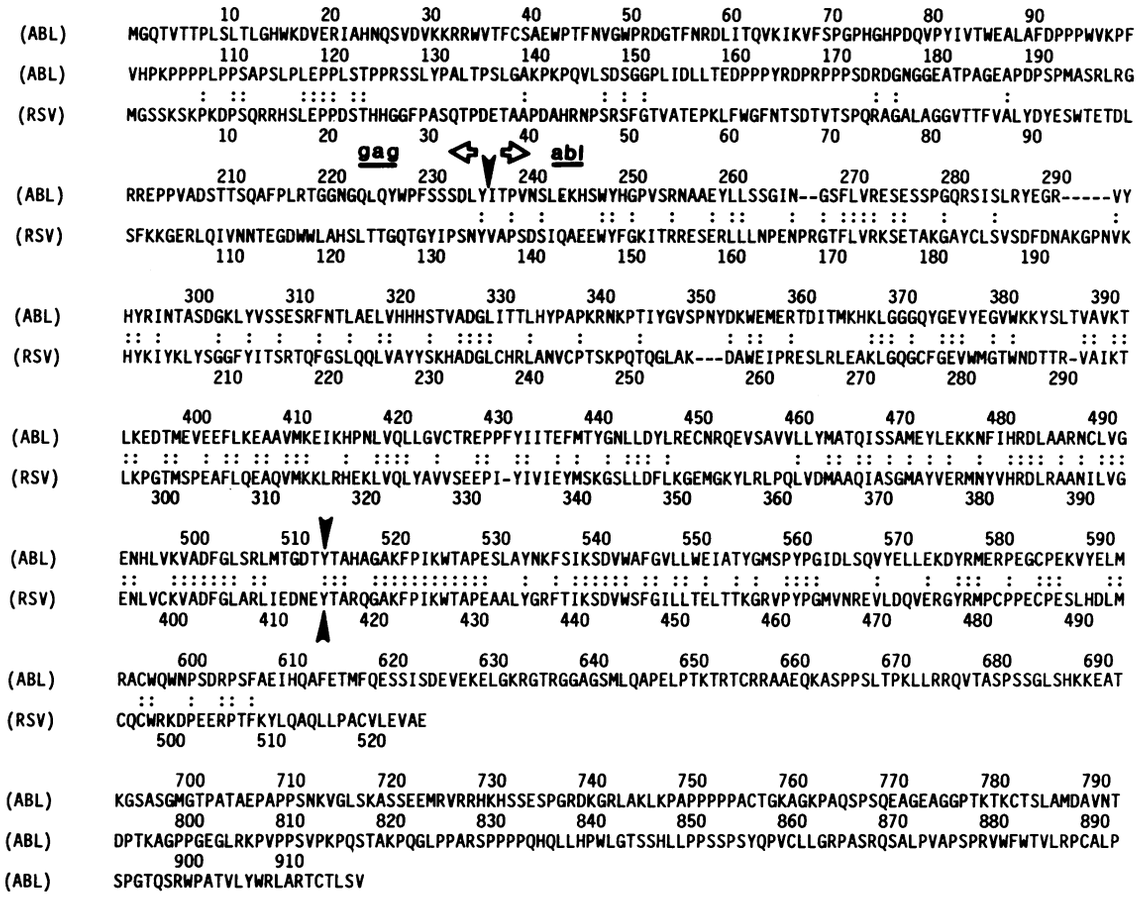

FIG. 3. Similarities between the deduced amino acid sequences of v-*abl* and v-*src* gene products. The two amino acid sequences are aligned to give maximal homology. The lower line gives the putative amino acid sequence of the RSV protein pp60^{src} (22) and the upper line gives the putative amino acid sequence of p120$^{gag-abl}$. The two amino acid sequences are aligned to give the maximal homology. For this purpose the computer methods described by Wilbur and Lipman (26) were used. A *K*-tuple size of 1, window size of 20, and gap penalty of 3 were used in this analysis. Dots indicate identical amino acids. The phosphate acceptor tyrosine residue is indicated by an arrow. A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr.

of a large open reading frame preceding the amino terminus of p15 and led to the speculation that such a precursor could be synthesized from a processed *gag* mRNA (12, 20). The sequence data presented here for the A-MuLV genome show that the viral genome has suffered a point mutation at position 1,054, creating a terminator codon at this position just five amino acids preceding the amino terminus of p15. It is, therefore, possible that A-MuLV transforming protein is not synthesized via a precursor polypeptide as is the case with the M-MuLV *gag* polyprotein.

It is also interesting to note that the viral genome contains an additional stretch of 802 bp of cell-derived sequences downstream from the terminator codon. The role of these sequences in the virus-induced transformation is unclear. However, this stretch of DNA appears to have a second open reading frame starting with an ATG at position 4,153–4,155 and ending with a TAG at position 4,642–4,644. This stretch of sequences could code for a protein of 163 amino acids with a molecular mass of 18,000 daltons. It is possible that the two reading frames arose as a result of point mutations or small deletions in the cellular insertion sequence creating a terminator codon at position 3,821–3,823, which otherwise could have been a contiguous reading frame in the c-*abl* encoded mRNA.

The finding that there is a considerable homology between the predicted sequences of p120 and other tyrosine-specific kinases is very striking. The optimal alignment of amino acid sequences with pp60^{src} revealed extensive amino acid homology with the carboxyl terminus of pp60^{src}. A single site for tyrosine phosphorylation has been detected within pp60^{src} (29). This residue has been located precisely at position 416 of the sequence shown in Fig. 3 for the Prague strain of RSV (22). The location of the phosphate acceptor tyrosine residue in pp60^{src} is homologous with the position of tyrosine residue at position 2,606–2,608 (marked with arrows in Figs. 1 and 3). These observations further support the argument that different retroviral transforming genes, all of which encode functionally related proteins with associated tyrosine-specific kinase activities, are derived from divergent members of the same protooncogene family (24, 25). The more extensive divergence of the 5′ ends of the protein molecule is puzzling. It is possible that this divergence indicates that this half of the molecule is relatively unimportant, allowing continued accumulation of point mutations. Alternative mechanisms such as exon shuffling and homologous recombination could also have contributed to this divergence.

Though v-*src* and v-*abl* gene products show extensive homology, they appear to have completely different cellular targets for transformation. The reason for this is unclear at the present time. However, examination of the two protein sequences (Fig. 3) reveals that pp60^{src} has an amino terminus 136 amino acids longer than the v-*abl* gene product. On the other hand, v-*abl* has a carboxyl terminus that is 294 amino acids longer

than that of pp60^{src}. It is possible these structural differences play a critical role in determining the tissue specificity of these two viruses.

Our earlier studies (7) had indicated that excision of the carboxyl terminus of the proviral genome deleting approximately 40% of the acquired cellular sequences does not affect the fibroblast-transforming activity of the proviral DNA molecule. Thus deletion of DNA sequences downstream from the *Sal* I restriction site at position 3,517 did not alter the transforming activity, whereas deletion of sequences upstream abolished such activity. Examination of the sequences presented in Figs. 1 and 3 reveals that in fact this stretch of proviral genome codes for the region of *abl* protein that has no counterpart in pp60^{src} and is therefore not required for fibroblast-transforming activity.

Like all other transforming genes of retroviruses, DNA sequences homologous to *abl* (termed c-*abl*) are found in normal mouse DNA (5, 15). Thus the viral oncogene represents a transduced cellular gene. The c-*abl* gene appears to be functionally active in several lymphoid organs, including thymus, coding for a protein product of 150,000 daltons (30, 31). Mouse thymus also synthesizes mRNA species 5.5 and 6.5 kbp long that cross-hybridize with v-*abl* sequences (our unpublished observations). Comparison of the size of the c-*abl*- and v-*abl*-encoded proteins shows that the c-*abl* gene codes for a protein approximately 60,000–70,000 daltons longer than that encoded by the v-*abl* gene sequence in the present studies. It, therefore, appears that extensive deletions of amino- or carboxyl-terminal sequences occurred during the recombinational process. Loss of amino- or carboxyl-terminal sequences during recombination between helper viral and cellular protooncogene sequences appears to be the rule rather than an exception. Thus, sequence analysis has revealed that amino-terminal deletions occurred during the generation of Moloney MuSV (16, 19, 20), simian sarcoma virus (17, 32), and myelocytomatosis (MC29) virus (33). Similarly, a loss of carboxyl-terminal sequences appears to have occurred during the generation of avian myeloblastosis virus (34, 35). In the case of Harvey MuSV (36), Kirsten MuSV (37), and BALB MuSV (our unpublished results), which appear to have acquired the entire coding region, the oncogenes appear to have undergone specific point mutations that play a crucial role in the transformation process (38, 39). Thus it appears that protooncogenes and viral oncogenes code for proteins that are structurally different. The availability of molecular clones of the protooncogenes makes it possible to analyze whether structural changes such as deletions and mutations in these genes lead to their biological activation.

We thank David Lipman, Richard Feldmann, and Sushil Devare for help with computer analysis and Stuart Aaronson for helpful advice and support.

1. Abelson, H. T. & Rabstein, L. S. (1970) *Cancer Res.* **3**, 2208–2212.
2. Potter, M., Sklar, M. D. & Rowe, W. P. (1973) *Science* **182**, 592–594.
3. Premkumar, E., Potter, M., Singer, P. A. & Sklar, M. D. (1975) *Cell* **6**, 149–159.
4. Rosenberg, N. & Baltimore, D. (1980) in *Viral Oncology*, ed. Klein, G. (Raven, New York), pp. 187–203.
5. Shields, A., Goff, S., Paskind, M., Otto, G. & Baltimore, D. (1979) *Cell* **18**, 955–962.
6. Srinivasan, A., Reddy, E. P. & Aaronson, S. A. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 2077–2081.
7. Srinivasan, A., Dunn, C. Y., Yuasa, Y., Devare, S. G., Reddy, E. P. & Aaronson, S. A. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 5508–5512.
8. Rosenberg, N. & Witte, O. (1980) *J. Virol.* **33**, 340–348.
9. Witte, O. N., Rosenberg, N., Paskind, M., Shields, A. & Baltimore, D. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 2488–2492.
10. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
11. Roychoudhury, R. & Wu, R. (1980) *Methods Enzymol.* **65**, 43–62.
12. Shinnick, T. M., Lerner, R. A. & Sutcliffe, J. G. (1981) *Nature (London)* **293**, 543–548.
13. Shimotohno, K., Mizutani, S. & Temin, H. (1980) *Nature (London)* **285**, 550–554.
14. Dhar, R., McClements, W. L., Enquist, L. W. & Vande Woude, G. F. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3937–3941.
15. Goff, S. P. & Baltimore, D. (1982) in *Advances in Viral Oncology*, ed. Klein, G. (Raven, New York), Vol. 1, pp. 127–139.
16. Van Beveren, C., Van Straaten, F., Galleshaw, J. A. & Verma, I. M. (1981) *Cell* **27**, 97–108.
17. Josephs, S. F., Dalla Favera, R., Gelman, E. P., Gallo, R. C. & Wong-Staal, F. (1983) *Science* **219**, 503–505.
18. Reddy, E. P., Reynolds, R. K., Watson, D. K., Schultz, R. A., Lautenberger, J. A. & Papas, T. S. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 2500–2504.
19. Reddy, E. P., Smith, M. J., Canaani, E., Robbins, K. C., Tronick, S. R., Zain, S. & Aaronson, S. A. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 5234–5238.
20. Reddy, E. P., Smith, M. J. & Aaronson, S. A. (1981) *Science* **214**, 445–450.
21. Bister, K. & Duesberg, P. H. (1982) in *Advances in Viral Oncology*, ed. Klein, G. (Raven, New York), Vol. 1, pp. 3–43.
22. Schwartz, D., Tizard, R. & Gilbert, W. (1982) in *Molecular Biology of Tumor Viruses: RNA Tumor Viruses*, eds. Weiss, R., Teich, N., Varmus, H. & Coffin, J. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 1340–1348.
23. Kitamura, N., Kitamura, A., Toyoshima, K., Hirayama, Y. & Yoshida, M. (1982) *Nature (London)* **297**, 205–208.
24. Hampe, A., Laprerotte, I., Gailber, F., Fedele, L. A. & Sherr, C. J. (1982) *Cell* **30**, 775–785.
25. Shibuya, M. & Hanafusa, H. (1982) *Cell* **30**, 787–795.
26. Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 726–730.
27. Kloetzer, W. S., Maxwell, S. A. & Arlinghaus, R. B. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 412–416.
28. Naso, R. B., Arcement, L. J. & Arlinghaus, R. B. (1975) *Cell* **4**, 31–38.
29. Smart, J. E., Oppermann, H., Czernilofsky, A. P., Purchio, A. F., Erikson, R. L. & Bishop, J. M. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 6013–6017.
30. Witte, O. N., Rosenberg, N. & Baltimore, D. (1979) *Nature (London)* **281**, 396–398.
31. Ponticelli, A. S., Whitlock, C. A., Rosenberg, N. & Witte, O. N. *Cell* **29**, 953–960.
32. Devare, S. G., Reddy, E. P., Law, J. D., Robbins, K. C. & Aaronson, S. A. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 731–735.
33. Watson, D. K., Reddy, E. P., Duesberg, P. H. & Papas, T. S. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 2146–2150.
34. Rushlow, K. E., Lautenberger, J. A., Papas, T. S., Baluda, M. A., Perbal, B., Chirikjian, J. G. & Reddy, E. P. (1982) *Science* **216**, 1421–1423.
35. Klempnauer, K. H., Gonda, T. J. & Bishop, J. M. (1982) *Cell* **31**, 453–463.
36. Dhar, R., Ellis, R. W., Shih, T. Y., Oroszlan, S., Shapiro, B., Maizel, J., Lowy, D. & Scolnick, E. (1982) *Science* **217**, 934–937.
37. Tsuchida, N., Ryder, T. & Ohtsubo, E. (1982) *Science* **217**, 937–938.
38. Tabin, C. J., Bradley, S. M., Bargmann, C. I., Weinberg, R. A., Papageorge, A. G., Scolnick, E. M., Dhar, R., Lowy, D. R. & Chang, E. H. (1982) *Nature (London)* **300**, 143–149.
39. Reddy, E. P., Reynolds, R. K., Santos, E. & Barbacid, M. (1982) *Nature (London)* **300**, 149–152.