Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 4718–4722, August 1983
Biochemistry

Envelope gene of the Friend spleen focus-forming virus: Deletion and insertions in 3′ gp70/p15E-encoding region have resulted in unique features in the primary structure of its protein product

(retrovirus/nucleotide sequence/acute erythroleukemia/polycythemia/membrane glycoprotein)

LINDA WOLFF*, EDWARD SCOLNICK†, AND SANDRA RUSCETTI*

*Laboratory of Genetics, National Cancer Institute, Bethesda, Maryland 20205; and †Merck Sharp & Dohme Research Laboratories, West Point, Pennsylvania 19486

Communicated by Peter K. Vogt, May 19, 1983

ABSTRACT **A nucleotide sequence was determined for the envelope (*env*) gene of the polycythemia-inducing strain of the acute leukemia-inducing Friend spleen focus-forming virus (SFFV) and from this the amino acid sequence of its gene product, gp52, was deduced. All major elements of the gene were found to be related to genes of other retroviruses that code for functional glycoproteins. Although the carboxyl terminus of gp52 is encoded by sequences highly related to sequences in its putative parent, ecotropic Friend murine leukemia virus, the majority of the protein (69%), including the amino terminus, is encoded by dualtropic virus-like sequences. Nucleotide sequence comparisons suggest that the nonecotropic region may be more closely related to the 5′ substitution in dualtropic mink cell focus-inducing viruses than it is to the 5′ end of xenotropic virus *env* genes. A large deletion and two unique insertions have been located in the *env* gene of polycythemia-inducing SFFV and may account for some of the unusual structural characteristics, aberrant processing, and pathogenic properties of gp52. As a consequence of the deletion, amino-terminal gp70 and carboxyl-terminal p15E-encoding sequences are juxtaposed and it appears that translation from the p15E region, 3′ to the deletion, continues in the standard reading frame used by other retroviruses. Insertions of six base pairs and one base pair at the very 3′ end of the gp52-encoding region result in a SFFV-unique amino acid sequence and a premature termination codon.**

The Friend spleen focus-forming virus (SFFV) is a replication-defective retrovirus that when introduced into susceptible strains of mice leads to the development of an acute erythroleukemia (1, 2). SFFV disease has an extremely short latency period; foci appear in 10 days and a grossly enlarged spleen in 14–21 days after intravenous injection of the virus (2, 3).

Despite the fact that the virus causes a rapid induction of erythroblastic proliferation, it does not contain a typical v-oncogene. Pathological properties of SFFV have been linked to the *env* gene, which codes for a 52,000-dalton glycoprotein, gp52 (4, 5). Our studies and those of others have shown that the *env* gene product differs from those of other retroviruses not only in its lower molecular weight but also in its failure to be released from cells and incorporated into virions (6, 7). gp52 contains exclusively high-mannose carbohydrates, and only a small proportion of the protein gets further processed to a 65,000-dalton cell surface form (6, 8, 9), suggesting that gp52 cannot be transported and processed like other retroviral envelope glycoproteins. In addition, our immunological studies of the protein have shown that gp52 contains a domain related to both xenotropic and dualtropic mink cell focus-inducing (MCF) viruses (6, 10). This is consistent with genetic studies which suggest that SFFV was derived from an infecting ecotropic virus that recombined with endogenous sequences like those of xenotropic and MCF viruses (11, 12). Our more recent experiments using monoclonal antibody and peptide mapping indicate that the domain of gp52 encoded by the endogenously acquired genetic region begins at the amino terminus of the protein (13).

We have determined the nucleotide sequence of the *env* gene of the molecularly cloned (14) polycythemia-inducing form of SFFV (designated $SFFV_P$) in order to obtain a better understanding of the genetic basis for some of the unusual properties and pathogenic effects of gp52.

MATERIALS AND METHODS

$SFFV_P$ Viral DNA. For DNA sequence analysis, subclones derived from molecularly cloned full-length genomic $SFFV_P$, designated 4-1a3 (14), were used. This virus is the Lilly–Steeves strain of $SFFV_P$, $SFFV_P$(LS), derived by passing the N-tropic Friend virus complex (1) together with a B-tropic helper virus (BALB–Tennant leukemia virus) in BALB/c mice (15).

Plasmid recombinants were extracted by using an alkaline extraction procedure (16) and purified in cesium chloride gradients. Fragments of DNA were prepared by using restriction endonucleases (New England BioLabs and Bethesda Research Laboratories) and purified by electrophoresis on 1–1.5% agarose gels or 4–8% polyacrylamide gels.

Nucleotide Sequencing. The sequencing analysis was performed according to the procedures of Maxam and Gilbert (17, 18). Restriction fragments were labeled at the 5′ end by using [γ-^{32}P]ATP (New England Nuclear) and polynucleotide kinase (P-L Biochemicals) or at the 3′ end by using cordycepin 5′-[α-^{32}P]triphosphate (New England Nuclear) and terminal deoxynucleotidyltransferase (New England Nuclear) (19). Sequence reactions used were the G, G+A, C+T, C, and A>C reactions. Sequences were visualized after electrophoresis on 20% polyacrylamide gels (0.4 mm × 40 cm), 12% polyacrylamide gels (0.4 mm × 80 cm), and 8% polyacrylamide gels (0.4 mm × 80 cm).

RESULTS AND DISCUSSION

Overall Structural Features of $SFFV_P$ *env* Gene and gp52. The nucleotide sequence of the $SFFV_P$ *env* gene, as well as the deduced amino acid sequence of the *env* gene product, gp52, are presented in Fig. 1. The 5′ end of the sequenced region contains two sequences, C-T-C-T-C-C-A-A-G (positions 6–14) and C-A-T-T-T-A-C-A-G (positions 17–25), that resemble the consensus sequence for splice acceptor sites (22) and are either

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.

Abbreviations: SFFV, spleen focus-forming virus; $SFFV_P$, polycythemia-inducing form of SFFV; MuLV, murine leukemia virus; MCF virus, mink cell focus-inducing virus; F, Friend; Mo, Moloney; bp, base pair; LTR, long terminal repeat.

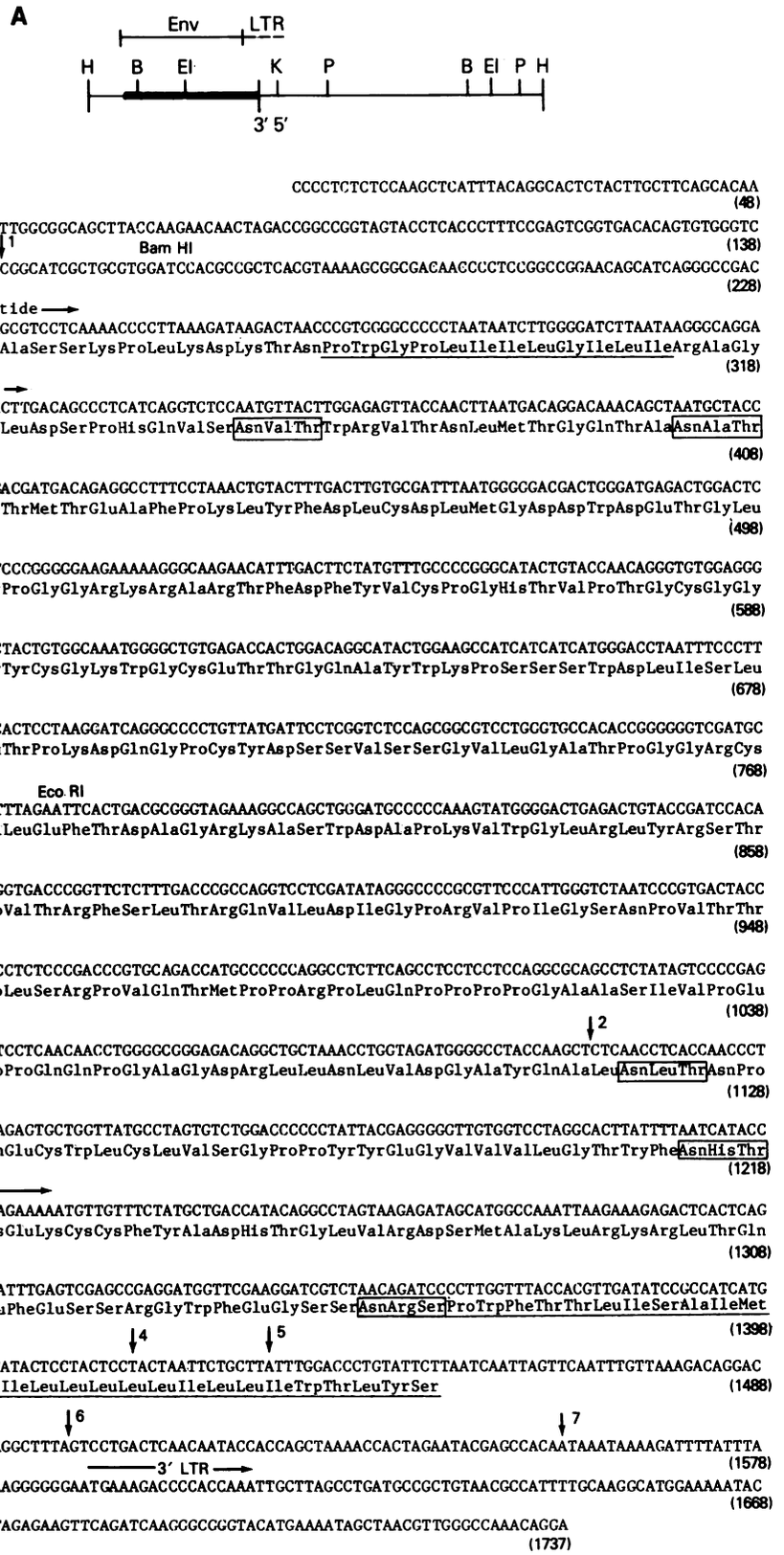

FIG. 1. (*A*) Restriction endonuclease map of the molecularly cloned $SFFV_P$ genome. The bold line represents the area of viral genome whose sequence was determined. B, *Bam*HI; EI, *Eco*RI; H, *Hind*III; K, *Kpn* I; P, *Pst* I. (*B*) Nucleotide sequence of the *env* gene of $SFFV_P$ and deduced amino acid sequence of gp52. Underlined areas designate uncharged regions of the amino acid sequence and boxes show potential glycosylation sites. Numbered arrows indicate: 1, location of a 74-base-pair (bp) deletion in the SFFV sequence as compared to the Moloney MCF (Mo-MCF) nucleotide sequence (20); 2, beginning of homology with Friend murine leukemia virus (F-MuLV) sequences (21); 3, beginning of a sequence encoding a p15E-related protein domain; 4, location of a 6-bp repeat insertion relative to other retroviral p15E-encoding sequences; 5, single base pair insertion in the SFFV genome relative to other retroviral p15E-encoding sequences; 6, location in other viruses, but not $SFFV_P$, of sequences encoding the cleavage site for the R peptide; 7, where termination of gp52 would occur if the $SFFV_P$ sequence did not have the single base pair insertion at nucleotide 1,426. Numbers in parentheses on the left are amino acids and numbers in parentheses on the right are base pairs. LTR, long terminal repeat.

identical or nearly identical to sequences proposed by others to be the potential splice acceptor site for the *env* mRNA of Moloney murine leukemia virus (Mo-MuLV) (20, 23).

An open reading frame, which undoubtedly encodes gp52 and encompasses 409 amino acids, begins at an ATG codon at position 229 and extends to nucleotide 1,455. Although the amino terminus of gp52 has not been defined by amino acid sequence analysis, the sequence starting Val-Gln-Leu-Asp-Ser and extending for 25 residues aligns, in 22 out of 25 positions, with the primary amino-terminal sequence determined for Rauscher MCF gp69 (24). The 32-amino-acid sequence that precedes this valine has characteristics in common with other amino-terminal peptides of membrane and secreted proteins (23, 25, 26) and probably represents a leader peptide. The nucleotide sequence encoding the gp52 leader peptide more closely resembles the sequence encoding the Mo-MCF *env* leader peptide (20) than

the corresponding F-MuLV sequence (21), suggesting that the nucleotide sequence encoding the leader peptide is a part of the major region of substitution acquired in the generation SFFV (see below). However, the SFFV sequence has a 74-bp deletion relative to the Mo-MCF sequence in the region preceding the ATG initiator codon (Fig. 1).

The gp52-encoding sequence consists of several distinct regions. The 5′ region contains sequences analogous to those that encode the amino terminus of MCF and xenotropic viral gp70s (20, 27). Downstream is a short sequence that is homologous with a small segment of F-MuLV gp70-encoding sequences (21) followed by a region corresponding to the 3′ portion of p15E-encoding sequences. It appears that the p15E-like region in SFFV is translated in the same reading frame as p15E sequences of other retroviruses and the translated product becomes part of the carboxyl terminus of gp52. The overall structure of the SFFV envelope glycoprotein, as deduced from the nucleotide sequence, is depicted in Fig. 4. gp52 resembles a retroviral envelope precursor protein with a large deletion in the middle and contains essentially the amino terminus of gp70 and the carboxyl terminus of p15E.

Downstream from the 3′ end of the open reading frame in the SFFV genome is the 3′ LTR, beginning at nucleotide 1,506. For the part of the SFFV LTR whose sequence we have determined, there is 95% homology with the LTR sequence of another cloned isolate of SFFV (28). The 41-bp sequence that precedes the LTR in SFFV is identical to the sequence preceding the LTR in F-MuLV (21) and is within the region proposed to contain the origin of plus strand DNA synthesis (29, 30).

MCF- and Xenotropic-Virus-Related Sequences in the $SFFV_P$ *env* Gene. Because previously obtained oligonucleotide mapping data and the heteroduplex analyses have shown $SFFV_P$ to have a major substitution in the *env* gene relative to F-MuLV (12, 31, 32), a comparison of their *env* sequences was made. The nucleotide sequence encoding the first 262 amino acids of gp52 (greater than two-thirds of the protein) and the leader peptide shares little homology with sequences in F-MuLV gp70 (21), but there is 95% homology of $SFFV_P$ sequences in this region with sequences encoding the amino terminus of Mo-MCF gp70 (20). In addition, there is extensive homology with sequences recently published by Repaske *et al.* (27) of the 5′ end of the *env* gene of NFS-th-1 xenotropic virus. In the latter case, 91% of the $SFFV_P$ sequence, from the ATG initiator codon to the *Eco*RI site, is present in the xenotropic virus *env* gene sequence. This extensive homology between SFFV and both xenotropic and MCF viral sequences confirms earlier hybridization experiments (11). However, our data indicate that the 5′ *env* gene sequences of SFFV and MCF viruses are more related to each other than to xenotropic sequences. Perhaps this region in SFFV, like that in MCF viruses, is derived from endogenous sequences that are not a part of xenotropic viruses (33). This analysis unambiguously clarifies that the region of substitution in the SFFV *env* gene includes sequences encoding the amino terminus of gp52 and confirms our earlier studies using monoclonal antibody to gp52 and peptide mapping (13).

An amino acid comparison of the substituted region of gp52 (which makes up 69% of the primary sequence) with analogous regions of Mo-MCF and NFS-th-1 xenotropic virus gp70s is presented in Fig. 2 and shows that a remarkable conservation of sequences exists among these proteins. However, SFFV gp52 is more closely related to Mo-MCF gp70 than to the xenotropic gp70, because 240 (92%) of the 262 amino acids present in this region of gp52 are also present in the Mo-MCF protein, whereas only 139 of the first 156 amino acids (89%) are shared with the xenotropic protein. In addition, gp52, like MCF gp70, does not contain an additional xenotropic insertion of four amino acids after residue 55 (Fig. 2).

We observed that proteins in the $SFFV_P$/MCF/xenotropic group, when aligned with ecotropic proteins, have regions that are insertions relative to ecotropic proteins. These are designated by hatched areas in Fig. 2 and include a three amino acid residue insertion starting at residue 36 in all three proteins and a 15-residue insertion starting at residue 128 in MCF gp70 and SFFV gp52 (residue 132 in the xenotropic gp70). Because the only property unique to murine dualtropic and xenotropic viruses is their ability to infect heterologous cells, one might speculate that these regions are involved directly or indirectly in the interaction of the gp70 molecules with cell surface receptors on heterologous cells.

Effect of a Large *env* Gene Deletion. The SFFV genome has undergone extensive deletion, as indicated by its size and defectiveness in virus replication. The sequence analysis has made it possible to localize one large deletion within the *env* gene; it is at the site in the sequence indicated by arrow number 3 in Fig. 1. The deletion corresponds to a region of the F-MuLV envelope precursor protein ($Pr85^{env}$) that includes both the carboxyl-terminal domain of gp70 and the amino-terminal domain of p15E (see Fig. 4). Apparently, the cleavage site used in processing the precursor to gp70 and p15E is missing as a result of the deletion so that the amino-terminal gp70 domain and the carboxyl-terminal p15E domain remain covalently linked in a single protein.

Within the deleted region other retroviruses encode two hydrophobic segments and three glycosylation sites (20, 21, 23). The gp52 sequence retains four potential glycosylation sites [with the sequence Asn-X-Ser(Thr)] in the gp70-related domain and one in the p15E-related domain (Figs. 1 and 4). Earlier studies have shown that the apparent molecular weight of gp52 is decreased by 7,000 when the protein is labeled in the presence of tunicamycin, an inhibitor of glycosylation (34). Such a difference would account for four mannose-rich carbohydrate moieties on gp52. This is consistent with the sequence analysis if one excludes the Asn-Arg-Ser sequence in the p15E region, a site that does not appear to be glycosylated in ecotropic virus p15E (35). The deletion in the *env* gene of SFFV, if it included signals for glycoprotein processing, may offer an explanation as to why the gp52 is inefficiently processed to the plasma membrane form (6, 9).

p15E-Related Sequences. Following the deletion in the SFFV *env* gene are sequences that are highly related to those encoding the 3′ half of F-MuLV p15E. The deletion has occurred in such a way that the reading frame of these sequences in $SFFV_P$ is the same as that of sequences encoding p15E in F-MuLV and other viruses. The p15E-related domain of $SFFV_P$ was probably not identified in the past because immunological reagents were not available that would react with this portion of p15E (34).

Fig. 3 shows a comparison of the SFFV p15E amino acid sequence with the amino acid sequences in the same region of other retroviral p15Es, including those of ecotropic F-MuLV (21), Mo-MuLV (23), and Akv (36, 38). Mo-MCF p15E (20) is identical to Mo-MuLV p15E in this region and Akr MCF 247 (37) is the same as Akv except where indicated in the bottom line in Fig. 3. We have included this information in an attempt to show comparisons in general between ecotropic p15E sequences and p15E sequences of recombinant viruses.

The p15E sequence comparison in Fig. 3 shows that gp52 has 15 amino acids that are unique to this protein. Interestingly, at each of these locations the amino acids in the p15Es of all the other viruses, including the MCF viruses mentioned above, are identical with each other. Many of these amino acid changes

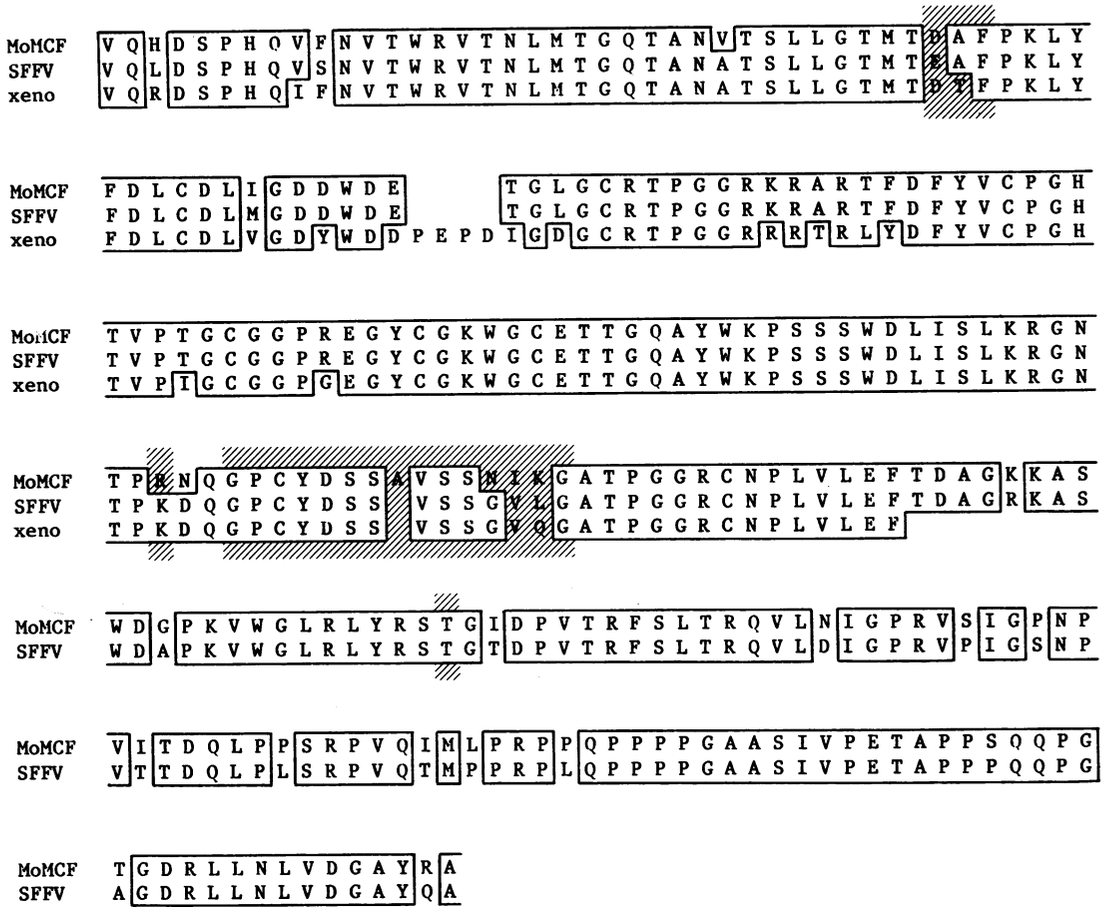

FIG. 2. Amino acid comparison of the substituted region of gp52 with analogous regions in the amino termini of Mo-MCF virus (20) and NFS-th-1 xenotropic virus (27) gp70s. The standard one-letter code for amino acids is used. Amino acids in the MCF and xenotropic proteins that are identical to amino acids in the SFFV protein sequence are enclosed in open boxes. Regions that are unique to MCF and xenotropic proteins and are insertions relative to ecotropic proteins are shown in the hatched areas.

occur in a segment of gp52 that is located at the carboxyl terminus. The nucleotide sequence in this segment of the coding region contains an extra 6-bp repeat (arrow 4, Fig. 1) followed shortly by a single base insertion (arrow 5, Fig. 1). The insertions cause gp52 to have a unique sequence that includes 8 amino acids not found in any of the other retrovirus p15Es. The single

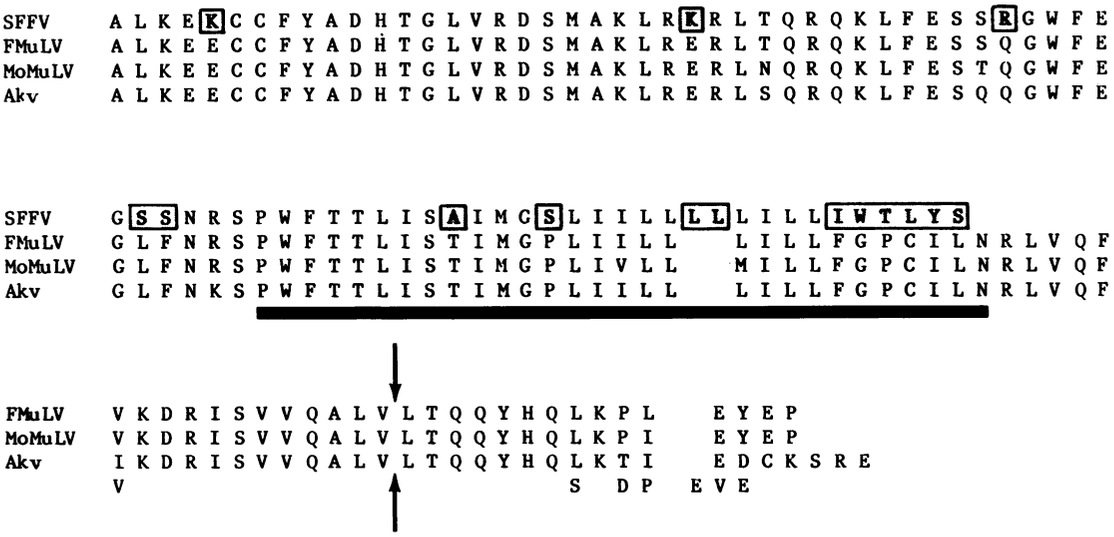

FIG. 3. Comparison of the $SFFV_P$ p15E-related amino acid sequence with the amino acid sequences of other retroviral p15Es in an analogous region, including those of F-MuLV (21), Mo-MuLV (23), and Akv (36). Amino acids that are unique to SFFV compared with the other viruses are enclosed in shaded boxes. The nonpolar region is underlined. Arrows indicate the cleavage site in Pr15E that generates p15E and the R peptide. Mo-MCF p15E is identical to Mo-MuLV p15E in this region (24) and Akr MCF 247 (37) is the same as Akv except where indicated in the bottom line below the Akv sequence.

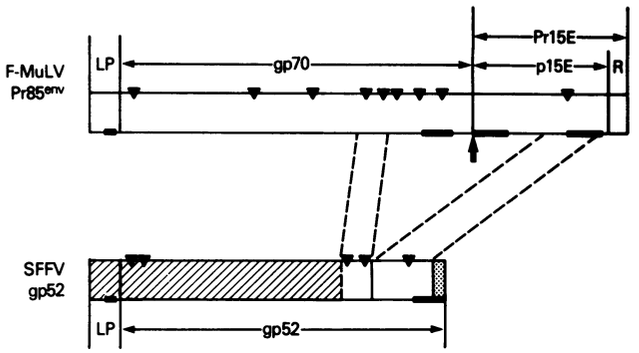

FIG. 4. Diagram comparing SFFV gp52 to F-MuLV $Pr85^{env}$. Hatched region is a domain of gp52 that is distantly related to F-MuLV. Open area in gp52 is encoded by sequences that are 93% homologous with F-MuLV sequences. Dotted area shows where gp52 has a unique p15E carboxyl terminus. LP, leader peptide; R, R peptide; ▼, potential glycosylation sites; heavy bars, uncharged regions.

base pair insertion also changes the reading frame in $SFFV_P$ so that the p15E sequence terminates early (Fig. 3), before the cleavage site (arrow in Fig. 3) that would normally generate the R peptide (39). Preliminary data from our laboratory show that an antiserum to the R peptide (39), which reacts with Pr15Es of F-MuLV and F-MCF virus, fails to precipitate gp52. Despite the change in amino acids at the end of gp52, the protein still retains a hydrophobic region which is as long as that found at the carboxyl terminus of other retroviral p15Es. The fact that gp52 has a carboxyl terminus like p15E may explain why it is so tightly bound to membranes and not secreted, as is gp70. We cannot determine, at present, the significance of the unique set of amino acids at the carboxyl terminus of gp52 and the lack of greater than 30 terminal amino acids found in other viral p15Es. However, we are considering their possible roles in determining the polycythemic feature of the $SFFV_P$-induced disease, target cell specificity, or disease latency.

Conclusions. $SFFV_P$ induces a unique murine erythroleukemia associated with an exceptionally short latency and a polycythemia. We performed a structural analysis of the *env* gene with the intent of revealing properties of the envelope glycoprotein, gp52, that could account for the unique disease characteristics. Some of the structural features of gp52, deduced from the analysis, are depicted in Fig. 4. It was found that major differences occur in two regions in this gene compared to other nondefective *env* genes. One difference is in the gp70/p15E junction region, where a large deletion has occurred, eliminating sequences encoding a natural cleavage site of the precursor protein, three glycosylation sites, and two hydrophobic domains. This deletion alone would affect the structure of gp52 in a major way. Other differences were observed in the p15E region, where the nucleotide sequence was shown to have two insertions; these changes should result in both a unique amino acid sequence and premature termination.

We thank David Linemeyer and John Menke for generously providing us with the molecular clones used in these studies and for their helpful advice. We are grateful to Lawrence Donehower for assisting us in the initiation of the sequence analysis and to Roy Repaske for providing us with the nucleotide sequence of the *env* gene of NFS-th-1 xenotropic virus prior to its publication. L.W. was supported by Public Health Service Fellowship F32 CA06662-02.

1. Friend, C. (1957) *J. Exp. Med.* **105,** 307–318.
2. Steeves, R. A. (1975) *J. Natl. Cancer Inst.* **54,** 289–294.
3. Axelrad, A. A. & Steeves, R. A. (1964) *Virology* **24,** 513–518.
4. Linemeyer, D. L., Ruscetti, S. K., Scolnick, E. M., Evans, L. H. & Duesberg, P. H. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 1401–1405.
5. Linemeyer, D. L., Menke, J. G., Ruscetti, S. K., Evans, L. H. & Scolnick, E. M. (1982) *J. Virol.* **43,** 223–233.
6. Ruscetti, S. K., Linemeyer, D., Feild, J., Troxler, D. & Scolnick, E. M. (1979) *J. Virol.* **30,** 787–798.
7. Dresler, S., Ruta, M., Murray, M. J. & Kabat, D. (1979) *J. Virol.* **30,** 564–575.
8. Ruscetti, S., Troxler, D., Linemeyer, D. & Scolnick, E. (1980) *J. Virol.* **33,** 140–151.
9. Ruta, M. & Kabat, D. (1980) *J. Virol.* **35,** 844–853.
10. Ruscetti, S., Linemeyer, D., Feild, J., Troxler, D. & Scolnick, E. (1978) *J. Exp. Med.* **148,** 654–663.
11. Troxler, D. H., Lowy, D., Howk, R., Young, H. & Scolnick, E. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 4671–4675.
12. Evans, L., Nunn, M., Duesberg, P. H., Troxler, D. & Scolnick, E. (1980) *Cold Spring Harbor Symp. Quant. Biol.* **44,** 823–835.
13. Wolff, L., Koller, R. & Ruscetti, S. (1982) *J. Virol.* **43,** 472–481.
14. Linemeyer, D. L., Ruscetti, S. K., Menke, J. G. & Scolnick, E. M. (1980) *J. Virol.* **35,** 710–721.
15. Lilly, F. & Steeves, R. A. (1973) *Virology* **55,** 363–370.
16. Birnboim, H. C. & Doly, J. (1979) *Nucleic Acids Res.* **7,** 1513–1523.
17. Maxam, A. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.
18. Maxam, A. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
19. Roychoudhury, R. & Wu, R. (1980) *Methods Enzymol.* **65,** 43–62.
20. Bosselman, R. A., van Straaten, F., van Beveren, C., Verma, I. M. & Vogt, M. (1982) *J. Virol.* **44,** 19–31.
21. Koch, W., Hunsmann, G. & Friedrich, R. (1983) *J. Virol.* **45,** 1–9.
22. Mount, S. M. (1982) *Nucleic Acids Res.* **10,** 459–472.
23. Shinnick, T. M., Lerner, R. A. & Sutcliffe, J. G. (1981) *Nature (London)* **293,** 543–548.
24. Schultz, A., Rein, A., Henderson, L. & Oroszlan, S. (1983) *J. Virol.* **45,** 995–1003.
25. Wicker, W. (1979) *Annu. Rev. Biochem.* **48,** 23–45.
26. Engelmann, D. M. & Steitz, T. A. (1981) *Cell* **23,** 411–422.
27. Repaske, R., O'Neill, R., Khan, A. S. & Martin, M. A. (1983) *J. Virol.* **46,** 204–211.
28. Clark, S. P. & Mak, T. W. (1982) *Nucleic Acids Res.* **10,** 3315–3330.
29. Gilboa, E., Mitra, S. W., Goff, S. & Baltimore, D. (1979) *Cell* **18,** 93–100.
30. Sutcliffe, J. G., Shinnick, T. M., Verma, I. M. & Lerner, R. A. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 3302–3306.
31. Evans, L. H., Duesberg, P. H., Troxler, D. H. & Scolnick, E. M. (1979) *J. Virol.* **31,** 133–146.
32. Bosselman, R. A., Van Griensven, L. D., Vogt, M. & Verma, I. M. (1980) *Virology* **102,** 234–239.
33. Chattopadhyay, S. K., Lander, M. R., Gupta, S., Rands, E. & Lowy, D. R. (1981) *Virology* **113,** 465–483.
34. Schultz, A. M., Ruscetti, S. K., Scolnick, E. M. & Oroszlan, S. (1980) *Virology* **107,** 537–542.
35. Pinter, A. & Fleissner, E. (1977) *Virology* **83,** 417–422.
36. Lenz, J., Crowther, R., Straceski, A. & Haseltine, W. (1982) *J. Virol.* **42,** 519–529.
37. Kelly, M., Holland, C. A., Lung, M. L., Chattopadhyay, S. K., Lowy, D. R. & Hopkins, N. H. (1983) *J. Virol.* **45,** 291–298.
38. Herr, W., Corbin, V. & Gilbert, W. (1982) *Nucleic Acids Res.* **10,** 6931–6944.
39. Green, N., Shinnick, T. M., Witte, O., Ponticelli, A., Sutcliffe, J. G. & Lerner, R. A. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 6023–6027.