Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 2500–2504, May 1983
Biochemistry

Nucleotide sequence analysis of the proviral genome of avian myelocytomatosis virus (MC29)*

(MC29 proviral genome/putative transforming protein)

E. PREMKUMAR REDDY†, R. K. REYNOLDS†, D. K. WATSON‡, R. A. SCHULTZ‡, J. LAUTENBERGER‡, AND T. S. PAPAS‡

Laboratories of †Cellular and Molecular Biology and of ‡Molecular Oncology, National Cancer Institute, Bethesda, Maryland 20205

Communicated by Max Tishler, January 20, 1983

ABSTRACT **The nucleotide sequence of the integrated proviral genome of avian myelocytomatosis virus (MC29) coding for *gag–myc* protein has been determined. By comparison of this nucleotide sequence with the helper virus as well as the c-*myc* region, it was possible to localize the junction points between helper viral and v-*myc* sequences. These studies demonstrate that (*i*) the large terminal repeat sequence of MC29 is very similar to that of Rous sarcoma virus, (*ii*) the viral genome has suffered extensive deletions in the *gag*, *pol*, and *env* genes, (*iii*) the *gag* region can code for p19, p10, and part of p27, (*iv*) the recombination between viral and cellular sequences occurred in the coding region of p27 such that the open reading frame extends for an additional stretch of 1,266 base pairs, resulting in a *gag–myc* hybrid protein, (*v*) the open reading frame terminated within the v-*myc* region 300 bases upstream of v-*myc*-helper viral junction, and (*vi*) the v-*myc* helper-viral junction at the 3′ end occurred in the middle of *env* gene, rendering it defective.**

Myelocytomatosis virus (MC29) is a replication-defective avian retrovirus that induces a broad spectrum of malignant diseases, including myelocytomas, renal and liver tumors, and, less typically, carcinomas, sarcomas, and erythroblastosis (1, 2). The same virus induces morphological transformation of fibroblasts, epithelial cells, and macrophages in culture (3–5). This virus arose by recombination of the nondefective helper virus (MC29-associated virus) and cellular sequences present within the normal chicken genome. These latter sequences termed "*myc*" appear to code for the transforming properties of the virus (6–8). The 5.7-kilobase (kb) RNA of MC29 has been shown to contain 1.6 kb of *myc* sequences that are flanked by partial Δ*gag* gene at the 5′ end and Δ*env* gene at the 3′ end (6). Nonproducer quail cell lines transformed by MC29 contain a 110,000-dalton protein with viral antigenic determinants (9). This protein appears to be a hybrid protein, the amino-terminal region of which is composed of helper virus *gag* gene products. The development of molecular cloning and DNA sequence analysis techniques has made the detailed analysis of the virus genome structure possible. In an attempt to better understand the structural organization and possible molecular mechanisms involved in transformation by MC29, we have undertaken primary DNA sequence analysis of the molecularly cloned integrated viral genome. Putative regulatory signals for transcription and translation of v-*myc* sequences have been identified. Sequence analysis has also demonstrated the occurrence of a long open reading frame within v-*myc* region that could code for the MC29 transforming protein.

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. The integrated proviral genome of MC29 was initially cloned in λgt WES·λB (10). In the present studies, two subclones of this DNA fragment were utilized for sequence analysis. A 1.4-kilobase-pair (kbp) *Xho* I fragment that contained the 5′ large terminal repeat (LTR) and *gag* sequences was subcloned in M13. The 2.9-kbp *Bam*HI fragment that contained the *gag*- and *myc*-specific sequence was subcloned in pBR322. The insert DNAs were purified by agarose gel electrophoresis and DEAE-cellulose (DE-52, Whatman) column chromatography after cleavage with appropriate restriction enzymes and were used in all subsequent analyses.

Nucleotide Sequence Analysis. Nucleotide sequence analysis of the 2.9-kbp *Bam*HI fragment was carried out by the method of Maxam and Gilbert (11). Appropriate restriction fragments were prepared and labeled at their 5′ ends by using [γ-^{32}P]ATP and polynucleotide kinase (P-L Biochemicals) as described by Maxam and Gilbert (11) or at their 3′ end by using cordycepin 5′-[α-^{32}P]triphosphate and terminal deoxynucleotidyl transferase (P-L Biochemicals) according to Roychoudhury and Wu (12). The nucleotide sequence was determined by the procedure of Maxam and Gilbert (11).

The 1.4-kbp *Xho* I fragment was subjected to sequence analysis by the dideoxy chain-terminator DNA sequence analysis procedure of Sanger *et al.* (13). The *Alu* I and *Sau*3A fragments of the 1.4-kbp DNA were ligated into M13 mp7 DNA restricted with the appropriate enzyme. Subclones prepared with M13 mp7 as the vector were identified as clear plaques and amplified. Single-stranded DNA was isolated and the sequence reactions were carried out as suggested by the cloning or sequence analysis kit suppliers (Bethesda Research Laboratories).

RESULTS

Strategy for Determining the Sequence of the MC29 Proviral Genome. The restriction map of molecularly cloned MC29 DNA was constructed by using both double-digestion analysis and the partial digestion technique of Smith and Birnstiel (14). Fig. 1 shows the restriction map of MC29 proviral DNA and the localization of its cell-derived sequences (v-*myc*). The sequence of both DNA strands was determined for >90% of the viral genome, and all restriction cleavage sites were confirmed by sequence analysis.

Sequence Organization of 5′ Noncoding Sequences. The nucleotide sequence of the MC29 proviral genome along with its

Abbreviations: MC29, avian myelocytomatosis virus strain MC29; kb, kilobase(s); kbp, kb pair; LTR, large terminal repeat; RSV, Rous sarcoma virus.

* Presented at the RNA Tumor Virus Meeting, Cold Spring Harbor, NY, May 26–30, 1982.

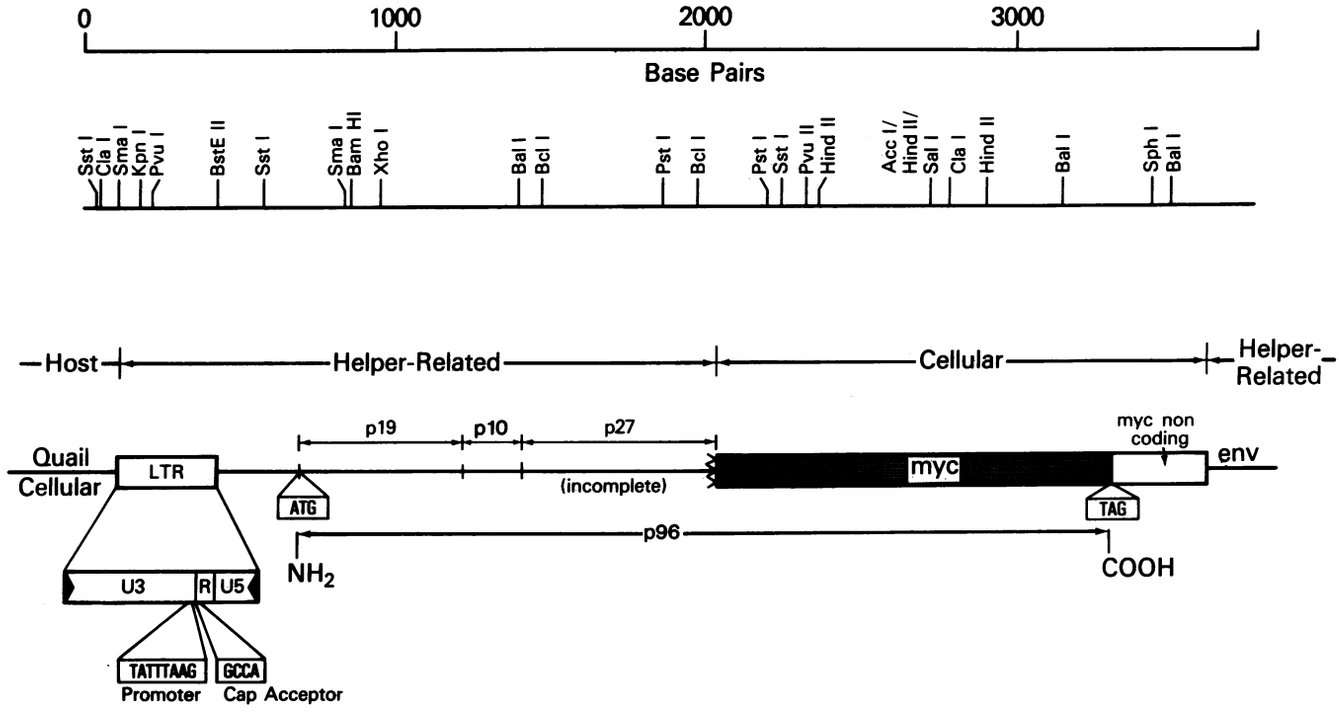

FIG. 1. Summary of the major structural features of the MC29 proviral genome. Restriction map and the important features of the MC29 viral genome, including the open reading frames, possible signals for the initiation of transcription and translation, are illustrated.

5′ noncoding and flanking quail cellular sequences is presented in Fig. 2. An important structural feature of the retroviral genome is the occurrence of two LTRs at both 5′ and 3′ ends of the proviral genome (for review, see ref. 15). Examination of the nucleotide sequence in Fig. 2 indicates the presence of such a LTR-like structure located at positions 114–426. Nucleotide sequence analysis of the LTRs of several avian and mammalian retroviruses has revealed that LTRs bear striking similarities to the terminal repeats of prokaryotic transposable elements and contain signals for the initiation and termination of transcription as well as mRNA capping (for review, see ref. 15). Thus, as with other known LTR sequences, we find the following features within the MC29 LTR: (*i*) inverted terminal repeats: the sequence T-G-T-A-G-T-C-T-T appears at the terminus of the LTR at positions 114–122 and in the inverted form at positions 418–426; (*ii*) a promoter-like sequence T-A-T-T-T-A-A-G was found at positions 298–305. This A+T-rich sequence precedes by 24 nucleotides the G-C-C-A sequence most likely to be the mRNA capping site; (*iii*) a polyadenylylation signal, A-A-T-A-A-A, was found in the LTR at positions 321–326. This signal preceded the dinucleotide C-A by 11 bases at position 338–339.

The DNA sequence of MC29 LTR region was compared to the Rous sarcoma virus (RSV) LTR sequence (16, 17) by the two-dimensional dot-matrix homology program (18). This analysis showed that there was 83% homology between the U_5 regions, whereas the homology between the U_3 regions was about 80%. This observation is consistent with earlier observations that U_3 regions of retroviral LTRs are most susceptible for sequence divergence (19, 20). Downstream from the 5′ LTR, a 19-base sequence complementary to the 3′ end of $tRNA^{Trp}$ (21) was localized at positions 428–446. Following this sequence was a stretch of 256 base pairs of noncoding sequences, followed by the *gag* gene coding sequences.

Sequence Organization of MC29 Transforming Gene. Downstream from the 5′ LTR sequences, we observed two ATG codons at positions 519–521 and 704–706. The first ATG codon is unlikely to function as the initiator codon as it was followed closely by in-phase termination codons. However, the second ATG codon at position 704–706 was followed by a large open reading frame of 2,625 bases ending with a TAG codon at position 3,329–3,331. We presumed that this sequence coded for the viral *gag–myc* hybrid protein. Fig. 2 shows the predicted amino acid sequence of this polypeptide. Earlier work on the biochemical and immunological analysis of the putative transforming protein of MC29 revealed that it is a 110,000-dalton polyprotein comprised of sequences derived both from the *gag* region and the cell-derived *myc* regions (9). To determine the exact point of recombinational event, we determined the sequence of the corresponding regions of the c-*myc* gene and also compared this with the known sequence for the RSV *gag* gene (17) (Fig. 3). Thus, it was possible to localize the junction points between the c-*myc* and helper viral sequences that are indicated in Figs. 2 and 3. It is interesting to note that there is a stretch of 10 bases in v-*myc* at the 5′ junction that did not show correspondence either to c-*myc* or RSV *gag* sequences. The origin of these sequences at this point is unclear. On the other hand, at the 3′ junction the v-*myc*, c-*myc*, and RSV *env* genes showed a sequence homology of five bases. Whether the occurrence of this homology is purely accidental or has a significance in the recombinational process between helper virus and host cellular sequences is speculative at this time.

The open reading frame shown in Fig. 2 could code for a polypeptide of 875 amino acids with a molecular mass of 96,000 daltons. This is in close agreement to the estimated size of 110,000 daltons for *gag–myc* hybrid protein synthesized by MC29-infected quail nonproducer cell lines. The difference in molecular masses could be due to post-translational modifications such as glycosylation and phosphorylation of the protein molecule. This protein consists of 450 amino acids derived from the amino terminus of *gag* region, followed by 422 amino acids that are specific for the v-*myc* region. Thus, the *gag* region contained the entire sequences of p19 and p10 but only the first 211 amino acids of p27 followed by the c-*myc*-derived sequences. The coding region terminated within the cell-derived *myc* sequences,

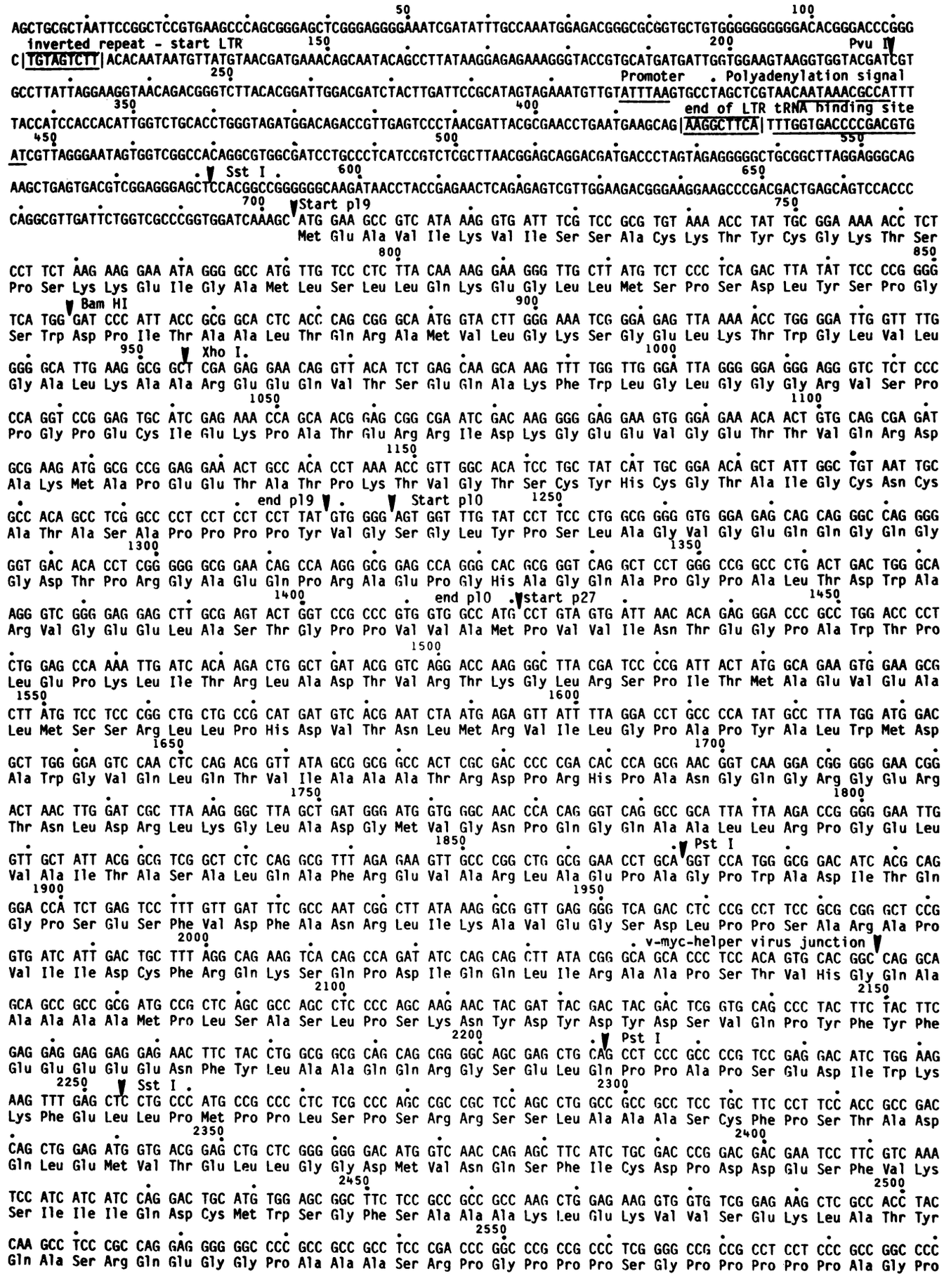
inverted repeat - start LTR
Pvu II
Promoter
Polyadenylation signal
end of LTR
tRNA binding site
Sst I
Start p19
Bam HI
Xho I.
end p19
Start p10
end p10
start p27
Pst I
v-myc-helper virus junction
Pst I
Sst I

FIG. 2. Nucleotide sequence of the MC29 proviral genome. The upper line shows the sequence proceeding in the 5′ to 3′ direction and has the same polarity on the MC29 genomic RNA. The amino acid sequence deduced from the open reading frames is given in the bottom line. The major structural features of the genome are indicated.

300 bases upstream from the v-*myc* helper-viral junction at the 3′ end.

Examination of the amino acid sequence of the putative transforming protein revealed that the carboxyl terminus of the v-*myc*-encoded protein was highly hydrophilic, containing a large number of glutamic acid, arginine, and lysine residues, whereas the amino terminus of the *myc*-encoded protein was more hydrophobic. This observation is further confirmed by the computer analysis of the v-*myc*-encoded protein by using the method of Hopp and Woods (22). The two regions are joined by a stretch of sequence that contained a large number of proline residues, indicating that these two regions of the polypeptide are linked by a highly flexible region. This structure is reminiscent of the hinge region present in immunoglobulin molecule at the junction of the variable and constant regions. It is interesting to note that these two biochemically distinct domains of the v-*myc* protein are derived from two different exons of the c-*myc* gene (7).

As mentioned earlier, the v-*myc* gene contains 300 base pairs of noncoding sequences at the 3′ end which did not contain any signals for termination of transcription or polyadenylylation of mRNAs. Comparison of our sequence presented here with that of c-*myc* (23) reveals that the polyadenylylation signal indeed occurs in the c-*myc* gene beyond the point of recombination. It is interesting to note that the recombination between the proviral and *myc* sequences occurs in the middle of envelope gene, deleting the first 75 codons, thus rendering it defective.

DISCUSSION

Nucleotide sequence analysis of the MC29 transforming region has revealed several important features of its molecular organization. Examination of the sequence data presented here reveals a single open reading frame on the viral RNA strand that could code for a protein of 875 amino acids with a molecular mass of 96,000 daltons.

Like many of the transforming retroviruses, MC29 appears to synthesize its transforming protein by means of a *gag–myc* polyprotein, the amino-terminal region of which is composed of helper virus *gag* gene products. In the case of MC29, the *gag–myc* hybrid protein contains the entire sequence of p19, p10, and the first 211 amino acids of p27 and 422 amino acids of v-*myc*. Thus, the transforming protein utilizes helper viral sequences for the initiation of its synthesis.

Like all other transforming genes of retroviruses, DNA sequences homologous to *myc* (termed c-*myc*) are found in normal chicken DNA (6–8). Thus, the viral oncogene represents a transduced cellular gene. Avian retroviruses of the MC29 group that carry v-*myc* genes cause an abnormally large variety of neoplasms, such as myelocytomas, endotheliomas, mesotheliomas, renal and hepatic carcinomas, and sarcomas (1, 2). MC29 virus also transforms both fibroblasts and macrophages in culture but not erythroblasts, even though p110 is synthesized in the infected erythroblasts (3–5). These experiments indicate that MC29 virus has certain target-cell specificity in spite of its wide spectrum of target cells. Furthermore, the disease spectrum can be restricted by creating deletions *in vitro* within the *onc* gene (24, 25). Such deletion mutants have been shown to transform fibroblasts but not macrophages *in vitro* (24, 25).

Lymphoid neoplasms are not usually caused by MC29 virus, yet the activation of c-*myc* by avian leukosis virus integration in the vicinity of the cellular gene is observed in a vast majority

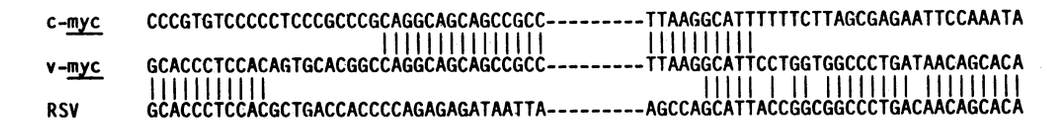

FIG. 3. The 5′ and 3′ recombination junctions between v-*myc* and c-*myc*. The 5′ and 3′ boundaries between helper virus and c-*myc* leading to the formation of MC29 virus are shown. The vertical lines indicate the areas of sequence homology. The RSV *gag* and *env* sequences are from Schwartz *et al.* (17). Note the presence of 10 bases in v-*myc* at the 5′ junction that has no apparent homology with RSV *gag* or c-*myc*.

of the bursal lymphomas induced by avian leukosis virus (26, 27). These observations led to the "downstream" promotion hypothesis that implicated the enhanced expression of c-*myc* protein in the transformation process induced by the avian leukosis virus. Recently, Collins and Groudine (28) and Dalla-Favera *et al.* (29) have reported an alternative mechanism for the increased transcription of c-*myc* gene in HL-60 cells. In these cells it appears that the increased transcription occurs as a result of amplification of c-*myc* genes. It is interesting to note that comparison of our v-*myc* sequence with that of c-*myc* (unpublished data) has revealed that v-*myc* in MC29 virus is an incomplete gene that has lost at least 300 base pairs of leader sequences at the amino-terminal region during the integration process. Therefore, it can be considered a deletion mutant of the normal c-*myc* gene and therefore has a decreased ability to transform lymphoid cells. The data on deletion mutants and revertants taken together suggest that mutations and deletions affect the tissue specificity of this *onc* gene.

Unlike all other known *onc* gene products, the p110 of MC29 is not a protein kinase and is not associated with cell membrane structure. Subcellular fractionation and immunofluorescence studies have shown that p110 is located in the nucleus of the transformed quail nonproducer cells (30, 31). Furthermore, purified p110 behaves as a DNA binding protein, suggesting that this protein may be involved in gene regulation. Because the putative amino acid sequence presented in Fig. 2 for v-*myc* region exhibits two biochemically distinct domains, it would be interesting to study the location of the sequences that are associated with the DNA binding properties of the molecule. Such studies may not only provide insights into the transformation process induced by MC29 but also shed light on the mechanisms involved in gene regulation of eukaryotic cells.

We are grateful to Richard Feldman for his help in computer analysis and Stuart Aaronson for helpful advice and support.

1. Beard, J. W. (1980) in *Viral Oncology*, ed. Klein, G. (Raven, New York), pp. 55–87.
2. Mladenov, Z., Heine, U., Beard, D. & Beard, J. W. (1967) *J. Natl. Cancer Inst.* **38**, 251–285.
3. Langlois, A. J., Sankaram, S., Hsiung, P.-H. L. & Beard, J. W. (1967) *J. Virol.* **1**, 1082–1084.
4. Bolognesi, D. P., Langlois, A. J., Sverak, L., Bonar, R. A. & Beard, J. S. (1968) *J. Virol.* **2**, 576–586.
5. Langlois, A. J., Fritz, R. B., Heine, U., Beard, D., Bolognesi, D. P. & Beard, J. W. (1969) *Cancer Res.* **29**, 2056–2074.
6. Mellon, P., Pawson, A., Bister, K., Martin, G. S. & Duesberg, P. H. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 5874–5878.
7. Robins, T., Bister, K., Garon, C., Papas, T. & Duesberg, P. (1982) *J. Virol.* **41**, 635–642.
8. Sheiness, D., Fanshier, L. & Bishop, J. M. (1978) *J. Virol.* **28**, 600–610.
9. Bister, K., Hayman, M. J. & Vogt, P. K. (1977) *Virology* **82**, 431–448.
10. Lautenberger, J. A., Schultz, R. A., Garon, C. F., Tsichlis, P. N. & Papas, T. S. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 1518–1522.
11. Maxam, A. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
12. Roychoudhury, R. & Wu, R. (1980) *Methods Enzymol.* **65**, 43–62.
13. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5468.
14. Smith, H. O. & Birnstiel, M. L. (1976) *Nucleic Acids Res.* **5**, 4537–4545.
15. Varmus, H. E. (1982) *Science* **216**, 812–820.
16. Swanstrum, R., Varmus, H. E. & Bishop, J. H. (1982) *J. Virol.* **41**, 535–541.
17. 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.
18. Maizel, J. V., Jr., & Lenk, R. P. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 7665–7669.
19. Rushlow, K., Lautenberger, J., Reddy, E. P. & Papas, T. (1982) *J. Virol.* **42**, 840–846.
20. Devare, S. G., Reddy, E. P., Law, J. D. & Aaronson, S. A. (1982) *J. Virol.* **42**, 1108–1113.
21. Harada, F., Sawyer, R. C. & Dahlberg, J. E. (1975) *J. Biol. Chem.* **250**, 3487–3497.
22. Hopp, T. P. & Woods, K. R. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 3824–3828.
23. Watson, D. K., Reddy, E. P., Duesberg, P. H. & Papas, T. S. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 2146–2150.
24. Bister, K., Ramsay, G. M. & Hayman, M. J. (1982) *J. Virol.* **41**, 754–766.
25. Ramsay, G. M. & Hayman, M. J. (1982) *J. Virol.* **41**, 745–753.
26. Hayward, W. S., Neel, B. G. & Astrin, S. M. (1981) *Nature (London)* **290**, 475–480.
27. Payne, G. S., Bishop, J. M. & Varmus, H. E. (1982) *Nature (London)* **295**, 209–214.
28. Collins, S. J. & Groudine, M. (1982) *Nature (London)* **298**, 679–682.
29. Dalla-Favera, R., Wong-Staal, F. & Gallo, R. C. (1982) *Nature (London)* **299**, 61–63.
30. Donner, P., Greiser-Wilkie, I. & Moelling, K. (1982) *Nature (London)* **296**, 262–266.
31. Abrams, H. D., Rohrschneider, L. R. & Eisenman, R. N. (1982) *Cell* **29**, 427–439.