Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 1079–1083, February 1984
Biochemistry

Nucleotide sequence analysis of the long terminal repeat of human T-cell leukemia virus type II

(retrovirus/DNA sequence/transcription)

KUNITADA SHIMOTOHNO*, DAVID W. GOLDE†, MASANAO MIWA*, TAKASHI SUGIMURA*, AND IRVIN S. Y. CHEN†

*National Cancer Center Research Institute, Chuo-ku, Tokyo 104, Japan; and †Division of Hematology-Oncology, Department of Medicine, University of California School of Medicine, Los Angeles, CA 90024

Contributed by Takashi Sugimura, November 14, 1983

ABSTRACT **The nucleotide sequence of the human T-cell leukemia virus type II (HTLV-II) long terminal repeat (LTR) and its surrounding regions were determined. Our results show the following structural features: (*i*) the LTR is 763 base pairs (bp) in length and consists of 314 ± 1 bp of region U3, 248 ± 1 bp of region R, and 201 bp of region U5; (*ii*) the terminal nucleotides in the LTR form an inverted repeat of T-G. . . . .C-A; (*iii*) 6-bp direct repeats of cellular sequences flanking the provirus were present; and (*iv*) the putative functional signals for initiation or termination of viral RNA synthesis were identified. Comparison of the HTLV-II LTR sequence with that previously published for adult T-cell leukemia virus (ATLV; HTLV-I) shows that the LTRs are distinct. Some small regions are conserved between HTLV-II and ATLV, involving sequences important for transcription and a sequence of 21 nucleotides repeated three times in U3. This 21-bp repeat may be important in regulating viral transcription in lymphoid cells.**

Human T-cell leukemia virus type II (HTLV-II) is associated with a patient (Mo) with a T-cell variant of hairy-cell leukemia (1–3). This virus is distinct from the more common isolates of human T-cell leukemia virus, designated human T-cell leukemia virus type I (HTLV-I) or adult T-cell leukemia virus (ATLV). The gag protein, p24, of HTLV-I and HTLV-II is only partially related immunologically (3). The viral genomes are also distinct as demonstrated by nucleic acid hybridization with molecular clones of HTLV-II (4).

HTLV-I or ATLV-associated malignancies are generally highly malignant and have a poor prognosis (5–7). By contrast, the T-cell variant of hairy-cell leukemia associated with HTLV-II is a relatively benign disease, and the patient is currently asymptomatic 8 yr after treatment by splenectomy, although he still has characteristic circulating hairy cells. As with HTLV-I (ATLV), HTLV-II is capable of transforming normal human peripheral blood lymphocytes *in vitro*, resulting in transformed cells of T-cell phenotype (8–11).

The mechanism of HTLV-associated leukemogenesis is unknown. Molecular cloning of the novel type II of HTLV (4), associated with a distinct disease, provides a unique opportunity to analyze and compare the molecular features that characterize this class of human virus and the associated leukemias. The sequences of the retrovirus long terminal repeat (LTR) are necessary for viral RNA transcription, reverse transcription, and integration (12, 13).

Also, the LTR has been directly implicated in the leukemogenic properties of some avian and mammalian retroviruses (14–18). We have determined the complete sequence of the HTLV-II LTR and flanking sequences. Most regions of the LTR show considerable differences in sequence from the published sequence of the ATLV LTR (19, 20), consistent with previous nucleic acid hybridization studies (4). Small stretches of sequence that include regions important for transcription and viral replication are highly conserved. Other short stretches of nucleotides of unknown function are conserved, in particular, a sequence of 21 nucleotides repeated three times in region U3.

MATERIALS AND METHODS

Materials. Restriction enzymes were purchased from Takara Shuzo (Kyoto, Japan), Bethesda Research Laboratories, or New England BioLabs. The large fragment of DNA polymerase I was from Bethesda Research Laboratories, and S1 nuclease and polynucleotide kinase were from Boehringer Mannheim. Radiolabeled nucleotides were from Amersham.

DNAs. Molecular cloning of a HTLV-II provirus has been described (4). The structures of a clone of λH6 carrying the provirus and cellular flanking sequences and of pH6-R0.8, a subclone of the 5′ LTR region of the provirus, are shown in Fig. 1.

Nucleotide Sequence Analysis. DNA fragments from these cloned DNAs digested with *Bam*HI, *Ava* I, *Taq* I, or *Hin*fI were labeled either at the 5′ ends or at the 3′ ends by polynucleotide kinase in the presence of [γ-^{32}P]ATP (3000 Ci/mmol; 1 Ci = 37 GBq) or DNA polymerase I in the presence of [α-^{32}P]dCTP or [α-^{32}P]dTTP (3000 Ci/mmol), respectively. Labeled fragments were further digested with an appropriate restriction enzyme to isolate fragments that were labeled at only one end. Nucleotide sequence determination was carried out by the method of Maxam and Gilbert (21). Analysis of the sequence homologies with the published sequence of ATLV was by the computer program of Korn *et al.* (22).

S1 Nuclease Mapping. The procedure followed was the method of Berk and Sharp (23). pH6-R0.8 DNA was digested with *Bam*HI followed by bacterial alkaline phosphatase; 5′ ends of the DNA fragments were labeled with [γ-^{32}P]ATP (3000 Ci/mmol) by polynucleotide kinase. The DNAs were further digested with *Taq* I. A DNA fragment corresponding to the region from nucleotide 48 to 363 in the LTR was isolated by acrylamide gel electrophoresis. The 5′-end-labeled DNA (10^5 cpm) was mixed with about 10 μg of total RNA from patient-derived Mo cells, which produce HTLV-II (2), and hybridized at 50°C overnight. Then the incubation mixture was digested with different amounts of S1 nuclease (50–500 units) and electrophoresed through a 10% acrylamide

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: HTLV-II, human T-cell leukemia virus type II; ATLV, adult T-cell leukemia virus (HTLV-I); kbp, kilobase pair(s); LTR, long terminal repeat.

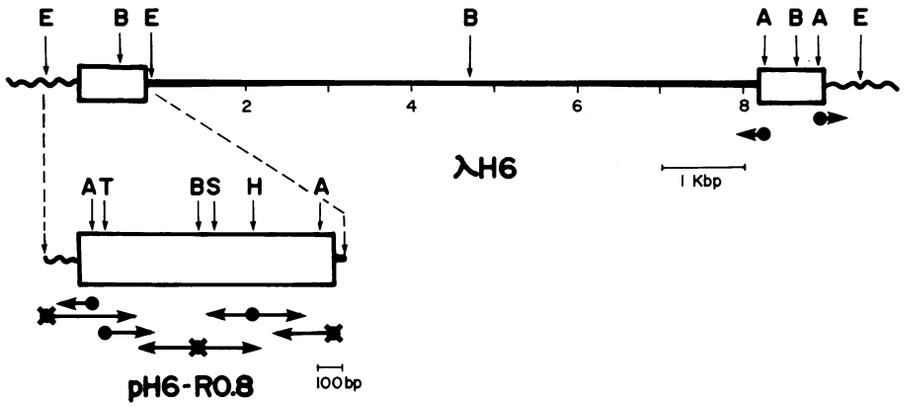

FIG. 1. Structure of the cloned HTLV-II provirus and sequence strategies. The HTLV-II provirus and its flanking cellular sequence in λH6 is shown. Wavy lines indicate the cellular sequence, boxes indicate the LTRs. pH6-R0.8 is the sublcone of the 5′ LTR region and some cellular flanking sequences. The nucleotide sequence was determined by using end-labeled DNA fragments as indicated: ●, 3′-end-labeled; ✖, 5′-end-labeled. Arrows and their length indicate the direction and size of regions whose sequence was determined. E, *Eco*RI; B, *Bam*HI; A, *Ava* I; T, *Taq* I; H, *Hin*fI; S, *Sac* I.

gel. Separately, the 5′-end-labeled DNA fragment was chemically treated for sequence assay and loaded onto the same gel to determine the size of the nuclease S1-resistant DNA fragments.

RESULTS

Nucleotide Sequence of LTR. λH6 represents a HTLV-II provirus that was molecularly cloned from DNA of a cell line (Mo) derived from the patient with a T-cell variant of hairy-cell leukemia (1, 4). The LTRs were identified by hybridization with a HTLV-II cDNA clone that represents the 3′ end of HTLV-II mRNA. The proviral DNA fragments used for sequence analysis are shown in Fig. 1. Restriction enzyme analysis demonstrated that the *Eco*RI fragment at the 5′ end of λH6 contains the entire 5′ LTR. The nucleic acid sequence surrounding restriction enzyme sites shown in Fig. 1 was determined.

The junction between the 5′ end of the LTR and flanking cellular sequences was ascertained by partial sequence determination of the 3′ LTR using the 3.5-kilobase pair (kbp) *Bam*HI fragment. The first nucleotide at which the 5′ LTR and 3′ LTR are identical was designated nucleotide 1. The 3′ end of the LTR was identified similarly by comparison with sequences of a 3.0-kbp *Bam*HI fragment that includes part of

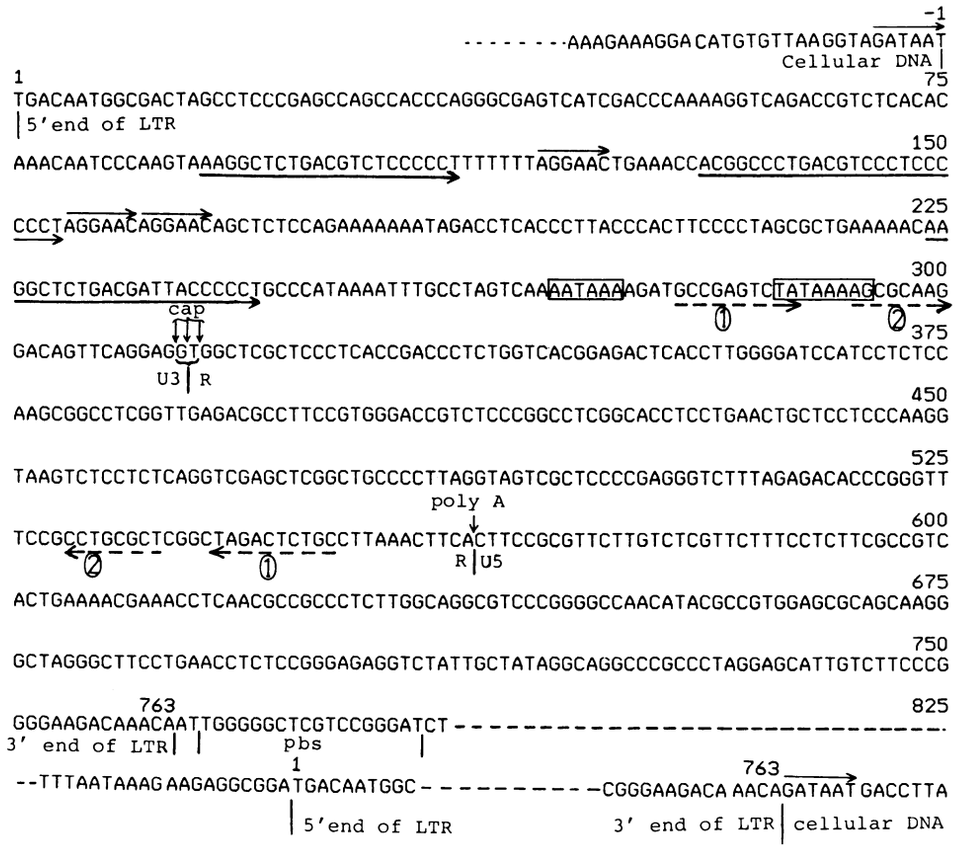

FIG. 2. Nucleotide sequence of HTLV-II LTR. →, Direct repeats; - - →, inverted repeats. The numbers underneath the dashed arrows indicate related inverted sequences. The putative polyadenylylation signal and promoter sequence are boxed. The cap site and polyadenylylation site are indicated with vertical arrows.

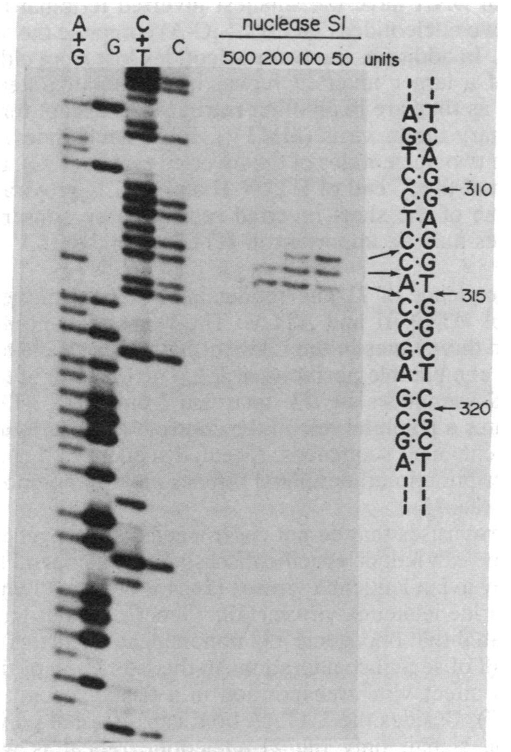

FIG. 3. Identification of the cap site of viral RNA. The 5′-end-labeled DNA fragment that remained after digestion with various amounts of S1 nuclease (500, 200, 100, or 50 units) was hybridized with Mo RNA and then subjected to 10% acrylamide gel electrophoresis. The chemically treated DNA fragment for nucleotide sequence analysis was also electrophoresed (left 4 lanes). The nucleotide sequence is shown. The 5′ ends of viral RNA determined by this procedure are shown.

the 3′ LTR and some cellular flanking sequences. The LTR of HTLV-II is 763 nucleotides in size. The complete nucleotide sequence is shown in Fig. 2.

Structural Features of the HTLV-II LTR and Flanking Sequences. As with all of the other retroviruses, the proviral DNA is terminated by the sequence T-G. . . . .C-A (12). The unintegrated viral DNA of retroviruses differs from proviral DNA by the presence of two additional terminal nucleotides at either end that are lost upon integration (12). These nucleotides usually form part of an inverted repeat of from 5 to 14 bp at the ends of unintegrated linear viral DNA. In HTLV-II these nucleotides are deduced to be G-A. . . . .A-T, but are not part of an inverted repeat. The only terminal nucleotides in HTLV-II proviral DNA that form an inverted repeat are the T-G. . . . .C-A nucleotides.

The R–U5 boundary (or polyadenylylation site of HTLV-II mRNA) was established by the sequence previously determined for the polyadenylylated 3′ end of HTLV-II mRNA (4). By analogy with other retroviruses, the R region ends with C-A. The junction of R–U5 was determined at nucleotide 562. The U3–R boundary (or cap site of HTLV-II mRNA) was established by determining the 5′ end of HTLV-II mRNA by the S1 nuclease mapping technique as shown in Fig. 3. For HTLV-II, U3 is 314 ± 1 bp, R is 248 ± 1 bp, and U5 is 201 bp. These sizes are similar to the sizes of the respective regions of the ATLV LTR.

A sequence of 18 nucleotides homologous to the 3′ end of $tRNA^{Pro}$ is present 2 nucleotides downstream from the 3′ end of the 5′ LTR. Therefore, $tRNA^{Pro}$ is probably used as a primer for the (−)-strand DNA synthesis by HTLV-II. $tRNA^{Pro}$ also is used for DNA synthesis by ATLV and by some murine and avian retroviruses (24).

The primer for synthesis of (+)-strand viral DNA is characterized by a purine-rich sequence located just upstream of the 3′ LTR. Although there is not a complete stretch of purine residues as found in avian or murine retroviruses, 15 out of 17 nucleotides are purines. Two pyrimidine bases in the 17-purine-rich sequences are located 4 and 15 nucleotides upstream from the 5′ end of the LTR. ATLV has a similar purine-rich sequence, and the pyrimidines are present at identical positions. Four of the 17 bases are different between HTLV-II and ATLV as shown in Fig. 4.

Integration of retroviruses results in a duplication of cellular sequences on either side of the provirus. The provirus of λH6 is characterized by a duplication of six bases of cellular DNA on either side.

Other important structural features of the LTR are indicated. The sequence T-A-T-A-A-A-A-G is present just upstream of the cap site (at 287–294) and is likely to be the "TATAA" box important for the initiation of transcription (25). The sequence A-A-T-A-A-A is present just upstream of the TATAA box (nucleotides 269–274) and is likely to be important in polyadenylylation of HTLV-II mRNA (26) (see below).

Comparison of the LTRs of HTLV-II and ALTV. The sequence of the HTLV-II LTR shows that the LTR of HTLV-II is quite different from that previously published for ATLV (19), consistent with the lack of detectable nucleic acid hybridization between the virus types in standard conditions (4). However, some important features of the LTRs are iden-

```
                 purine rich region | 5'end of LTR
           -20                      1    8           91                          111       131                 154
HTLV-II : ..TTTAATAAAGAAGAGGCGGATGACAATG...(83 bp)...AAGGCTCTGACGTCTCCCCCT..(19 bp)..ACGGCCCTGACGTC CC T C CCCCCT..
ATLV    : ..TTTAACGAAAAAGAGGCAGATGACAATG...(93 bp)...AAGGCTCTGACGTCTCCCCCC..(27 bp)..TAGGCCCTGACGTG   T  CCCCCT..
           -20                      1    8           102                         122       150                 170

                 224                245          269  274         282          296        301
HTLV-II : ..(69 bp)...AAGGCTCTGACGATTA C CCCCT..(23 bp)...AATAAA...(7 bp)...GAGTCTATAAAAGCG...(4 bp)..GACAGTTCAG
ATLV    : ..(79 bp)...CAGGCGTTGACGACAA  CCCCT..(36 bp)...AATAAA...(7 bp)...GAGTCTATAAAAGCG...(4 bp)..GACAGTTCAG
                 250                270          307  312         320          334        339

               322         427          439          763     766            783
HTLV-II : GAGGTGGCTCGC..(104 bp)..CCTCCTGAACTGC..(322 bp)..CA..(2 bp)..TGGGGGCTCGTCCGGGAT....
ATLV    : GAGGGGGCTCGC..( 85 bp)..CCTCCTGAACTGC..(294 bp)..CA..(2 bp)..TGGGGGCTCGTCCGGGAT....
               360         446          458          754     757            774
                                        3'end of LTR  |     primer binding site
```

FIG. 4. Nucleotide sequence homology between HLTV-II and ATLV in the LTR and its surrounding regions. The numbers indicate the position from the 5′ end of the LTR.

tical between HTLV-II and ATLV (Fig. 4). Twenty-one of 22 contiguous bases surrounding the cap site are identical in HTLV-II and ATLV. Fifteen contiguous nucleotides are identical around the TATAA box and A-A-T-A-A-A is present in the identical position of the ATLV LTR.

Most other sequences in region U3 were unrelated between HTLV-II and ATLV with a few notable exceptions. Other retroviruses have sequences in the U3 portion of the LTR, often present in more than one copy, which are necessary for efficient transcription. We identified a sequence of 21 nucleotides beginning at position 91. These nucleotides are repeated at position 131 with two mismatches and an insertion of three bases. A third repeat is present at position 224 with two mismatches and an insertion of one base. Twenty of 21 contiguous nucleotides of the upstream repeat are identical to a sequence at position 102 of the ATLV LTR. Although the authors did not mention this sequence in the discussion of the sequence of ATLV, by examining their published sequence we noted that this sequence is also repeated in the ATLV genome at position 150 with three mismatched nucleotides and at position 250 with six mismatched nucleotides. All three repeats of this sequence are present in comparable positions in both HTLV-II and ATLV. This 21-bp repeat is likely to be important for the control of HTLV-II and HTLV-I (ATLV) transcription.

A short sequence, A-G-G-A-A-C, is repeated in tandem between the second and third 21-bp repeat at position 155. The sequence G-A-A-G-C-C-A is repeated in a comparable position of the ATLV LTR (19). Eight nucleotides are identical between HTLV-II and ATLV at the exact 5′ end of the LTR including the dinucleotide T-G. Thirteen nucleotides are identical at position 427 in region R.

DISCUSSION

Two types of human T-cell leukemia virus have been identified. The viruses were associated with distinct diseases (1, 3, 5–7), and the viral genomes were shown to be distinct by restriction enzyme mapping and nucleic acid hybridization of molecular clones (4). Despite these differences, the viruses are related by some enzymological, serological, and biological properties (3, 11). The complete sequence of the HTLV-II LTR and flanking regions demonstrates that HTLV-II is related to HTLV-I (ATLV) at the molecular level.

Features of the LTR that were previously shown to distinguish the ATLV LTR from other retrovirus LTRs (19) are also conserved in HTLV-II. The R region is longer than that of most other retroviruses and similar in size to that reported for ATLV. Another unique feature of the HTLV-II and ATLV is the unusual position of the signal for polyadenylylation, A-A-U-A-A-A. As with ATLV, this sequence is located in the HTLV-II LTR upstream of the TATAA box about 290 nucleotides from the 3′ end of the mRNA. Usually this sequence is located 10–50 nucleotides from the 3′ end of a mRNA. Seiki *et al.* (20) proposed a secondary structure of the 3′ end of mRNA that would result in a closer proximity of the A-A-U-A-A-A sequence and the 3′ end of the mRNA. Our results also show a similar potential secondary structure. Sequences from nucleotide 279 to 288 and nucleotide 551 to 542 and sequences from nucleotide 293 to 300 and nucleotide 537 to 530 form inverted repeats, respectively (see Fig. 2).

Apart from those regions of the LTRs thought to be important for viral transcription and replication, most other regions of the LTRs of HTLV-II and ATLV show considerable differences in sequence consistent with previous nucleic acid hybridization studies (4). A few exceptions are notable. Recent evidence indicates that the inverted terminal sequences of viral DNA are recognized as a substrate for retrovirus integration (27). Among retroviruses, HTLV-II and ATLV proviral DNA have the smallest inverted terminal repeats. Only two nucleotides (T-G. . . . .C-A) comprise the inverted repeat. In addition, the two nucleotides lost from either end part of a larger inverted repeat in the unintegrated linear DNA, as they are in all other retroviruses except for mouse mammary tumor virus (MMTV). Eight nucleotides, including the two nucleotides of the inverted repeat (T-G), are conserved at the 5′ end of HTLV-II and ATLV proviral DNA. Because of the short inverted repeat, these conserved sequences may be important in HTLV-II and ATLV integration.

A sequence of 21 nucleotides is conserved in the U3 region of HTLV-II and ATLV. The same sequences are repeated three times in the LTRs of both viruses and are present in comparable positions of U3. The location of these repeated sequences in U3 upstream from the TATAA box indicates a potential role in the control of transcription similar to enhancer sequences, repeated sequences that regulate transcription in other animal viruses and in immunoglobulin genes (28–33).

Retroviruses that do not carry a cellular oncogene in their genome can induce specific diseases such as bursal lymphomas by avian leukemia viruses (14, 15) or T-cell lymphomas by murine leukemia viruses (18). Genetic studies have demonstrated that oncogenic and nononcogenic isolates differ as a result of sequence alterations in the U3 region of the LTR, which affect viral transcription in a tissue-specific manner (16, 17). Besides the TATAA box, cap site, and polyadenylylation signal, only the 21-nucleotide repeat is well conserved between ATLV and HTLV-II in regions of the LTR known to be important for transcription regulation. HTLV-II and HTLV-I (ATLV) appears to replicate only in T and B lymphocytes (5–11, 34). It is likely that these repeated sequences function as enhancer sequences, perhaps by conferring lymphoid specificity to the LTR.

We thank J. Gasson and J. McLaughlin for their assistance in this project. This work was partly supported by a Grant-in-Aid for Cancer Research from the Ministry of Education, Science and Culture and from the Ministry of Health and Welfare in Japan; by Grants CA 30388, CA 32737, and CA 09297 awarded by the National Cancer Institute; and by Grant PF-2182 from the American Cancer Society.

1. Saxon, A., Stevens, R. H. & Golde, D. W. (1978) *Ann. Intern. Med.* **88,** 323–326.
2. Saxon, A., Stevens, R. H., Quan, S. G. & Golde, D. W. (1978) *J. Immunol.* **120,** 777–782.
3. Kalyanaraman, V. S., Sarngadharan, M. G., Robert-Guroff, M., Miyoshi, I., Blayney, D., Golde, D. & Gallo, R. C. (1982) *Science* **218,** 571–573.
4. Chen, I. S. Y., McLaughlin, J., Gasson, J. C., Clark, S. C. & Golde, D. W. (1983) *Nature (London)* **305,** 502–505.
5. Poiesz, B. J., Ruscetti, F. W., Gazdar, A. F., Bunn, P. A., Minna, J. D. & Gallo, R. C. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 7415–7419.
6. Hinuma, Y., Nagata, K., Hanaoka, M., Nakai, M., Matsumoto, T., Kinoshita, K.-I., Shirakawa, S. & Miyoshi, I. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 6476–6480.
7. Blattner, W. A., Kalyanaraman, V. S., Robert-Guroff, M., Lister, T. A., Galton, D. A. G., Sarin, P. S., Crawford, M. H., Catovsky, D., Greaves, M. & Gallo, R. C. (1982) *Int. J. Cancer* **30,** 257–264.
8. Miyoshi, I., Kubonishi, I., Yoshimoto, S., Akagi, T., Ohtsuki, Y., Shiraishi, Y., Nagata, K. & Hinuma, Y. (1981) *Nature (London)* **294,** 770–771.
9. Yamamoto, N., Okada, M., Koyanagi, Y., Kannagi, M. & Hinuma, Y. (1982) *Science* **217,** 737–739.
10. Popovic, M., Sarin, P. S., Robert-Guroff, M., Kalyanaraman, V. S., Mann, D., Minowada, J. & Gallo, R. C. (1983) *Science* **219,** 856–859.
11. Chen, I. S. Y., Quan, S. G. & Golde, D. W. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 7006–7009.
12. Temin, H. M. (1981) *Cell* **27,** 1–3.

13. Temin, H. M. (1982) *Cell* **28,** 3–5.
14. Payne, G. S., Courtneidge, S. A., Crittenden, L. B., Fadly, A. M., Bishop, J. M. & Varmus, H. E. (1981) *Cell* **23,** 311–322.
15. Hayward, W. S., Neel, B. G. & Astrin, S. M. (1981) *Nature (London)* **290,** 475–480.
16. DesGroseillers, L., Rassart, E. & Jolicoeur, P. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 4203–4207.
17. Robinson, H. L., Blais, B. M., Tsichlis, P. N. & Coffin, J. M. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 1225–1229.
18. Tsichlis, P. N., Strauss, P. G. & Hu, L. F. (1983) *Nature (London)* **302,** 445–449.
19. Seiki, M., Hattori, S. & Yoshida, M. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 6899–6902.
20. Seiki, M., Hattori, S., Hirayama, Y. & Yoshida, M. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3618–3622.
21. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
22. Korn, L. J., Queen, C. L. & Wegman, M. N. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 4401–4405.
23. Berk, A. J. & Sharp, P. A. (1977) *Cell* **12,** 721–732.
24. Coffin, J., (1982) in *RNA Tumor Viruses, Molecular Biology of Tumor Viruses*, eds. Weiss, R. A., Teich, N., Varmus, H. A. & Coffin, J. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 1109–1203.
25. Ziff, E. B. & Evans, R. M. (1978) *Cell* **15,** 1463–1475.
26. Proudfoot, N. J. & Brownlee, G. G. (1976) *Nature (London)* **263,** 211–214.
27. Panganiban, A. T. & Temin, H. M. (1983) *Nature (London)* **306,** 155–160.
28. Banerji, J., Rusconi, S. & Schaffner, W. (1981) *Cell* **27,** 299–308.
29. De Villiers, J. & Schaffner, W. (1981) *Nucleic Acids Res.* **9,** 6251–6264.
30. Levinson, B., Khoury, G., Vande Woude, G. & Gruss, P. (1982) *Nature (London)* **295,** 568–572.
31. Gillies, S. D., Morrison, S. L., Oi, V. T. & Tonegawa, S. (1983) *Cell* **33,** 717–728.
32. Banerji, J., Olson, L. & Schaffner, W. (1983) *Cell* **33,** 729–740.
33. Queen, C. & Baltimore, D. (1983) *Cell* **33,** 741–748.
34. Yamamoto, N., Matsumoto, T., Koyanagi, Y., Tanaka, Y. & Hinuma, Y. (1982) *Nature (London)* **299,** 367–369.