Nucleic acid structure and expression of the human AIDS/lymphadenopathy retrovirus

Mark A. Muesing, Douglas H. Smith, Cirilo D. Cabradilla*, Charles V. Benton†, Laurence A. Lasky‡ & Daniel J. Capon

Departments of Molecular Biology and ‡ Vaccine Development, Genentech, Inc., 460 Point San Bruno Boulevard, South San Francisco, California 94080, USA

* Center for Infectious Disease, Centers for Disease Control, Atlanta, Georgia 30333, USA

The 9,213-nucleotide structure of the AIDS/lymphadenopathy virus has been determined from molecular clones representing the integrated provirus and viral RNA. The sequence reveals that the virus is highly polymorphic and lacks significant nucleotide homology with type C retroviruses characterized previously. Together with an analysis of the two major viral subgenomic RNAs, these studies establish the coding frames for the gag, pol *and* env *genes and predict the expression of a novel gene at the 3′ end of the genome unrelated to the X genes of HTLV-I and -II.*

ACQUIRED immune deficiency syndrome (AIDS) is a novel, transmissible deficiency of cellular immunity characterized by opportunistic infections and certain malignancies, notably *Pneumocystic carinii* pneumonia and Kaposi's sarcoma, in patients without another recognized cause for contracting these rare diseases[1-3]. AIDS is manifested by a profound lymphopenia, a generalized cutaneous anergy and a markedly reduced proliferative response to mitogens, antigens and allogeneic cells, seeming to result from depletion of the $OKT4^+$ T-lymphocyte subset[4]. While humoral immunity is relatively unaffected, there is increasing evidence for a hyperactive B-cell proliferative response which may be related causally to the high incidence of B-lymphoma in AIDS patients[5,6]. In addition to the fully-developed syndrome, an epidemic of a related disease, AIDS-related complex (ARC), has appeared, characterized by generalized chronic lymphadenopathy. This syndrome shares many of the epidemiological features and immune abnormalities and often precedes the clinical manifestations of AIDS. The predominant risk groups for AIDS and ARC include homosexually active males, intravenous drug abusers, recipients of transfusions and blood products and the heterosexual partners and children of high-risk individuals, suggesting the involvement of an infectious agent transmitted through intimate contact or blood products.

Recent evidence has implicated strongly a novel lymphocytotropic retrovirus as the primary aetiological agent of AIDS and the AIDS-related complex. Lymphadenopathy-associated virus (LAV) was isolated initially from cultured lymph-node T cells of patients with lymphadenopathy and AIDS as well as an AIDS patient and an asymptomatic sibling, both with haemophilia B[7-9]. A similar virus, designated human T-lymphotrophic virus type III (HTLV-III), has been isolated from a large number of AIDS and ARC patient blood samples by co-cultivation with the permissive T-cell line H9 (refs 10, 11). LAV and HTLV-III, as well as related retroviruses isolated recently from AIDS patients[12,13], share several important characteristics. Viral replication occurs in the $OKT4^+$ T-lymphocyte population *in vivo* and *in vitro* and is associated with impaired proliferation and the appearance of cytopathic effects[8,10,14]. The virus has a Mg^{2+}-dependent reverse transcriptase, exhibits a dense cylindrical core morphology similar to type D retroviruses[8,13,15] and is recognized by antibodies found in the sera of virtually all AIDS and ARC patients[8,13,16-21].

As a first step towards characterizing the molecular biology of this virus, we have determined the entire nucleotide sequence for one of the integrated proviruses present in H9/HTLV-III cells and for a complete set of overlapping complementary DNAs representing the viral RNA of distinguishable isolate(s) also present in H9/HTLV-III cells. Our results establish that LAV/HTLV-III has no nucleotide homology with previously characterized animal and human retroviruses and that different virus isolates display significant genetic heterogeneity. Together with transcriptional mapping data, these studies provide a detailed picture of the structure and processing of the *gag*, *pol* and *env* gene products, provide evidence for a novel gene in the 3′ region of HTLV-III and predict a further gene product in the region between the *pol* and *env* genes.

Isolation of cDNA and provirus clones

Molecular clones of HTLV-III were identified initially from cDNA libraries representing cellular RNA of productively infected H9/HTLV-III cells, established by Popovic *et al.*[10]. The H9 human T-cell line is permissive for the continuous production of high titres of HTLV-III isolated from the cultured lymphocytes of AIDS and ARC patients and is significantly resistant to the cytopathic effects of these viruses. The virus produced by H9/HTLV-III retains its cytopathic activity against fresh normal human lymphocytes[10]. Total poly(A)$^+$RNA was prepared from H9/HTLV-III cells infected with pooled material from several different AIDS patients[10] and used to construct an oligo(dT)-primed cDNA library in the vector λgt10.

The strategy used to identify clones containing HTLV-III sequences was based on differential hybridization with cDNA probes prepared from poly(A)$^+$RNA of H9/HTLV-III cells and the uninfected CEM human T-lymphoblastoid cell line; ~0.2% of the clones in this library contained inserts hybridizing specifically with the H9/HTLV-III cDNA probe. Six of these clones were purified and their DNA inserts used as probes to classify 50 additional H9/HTLV-III-specific clones. Two distinct classes of clones were identified on the basis of their pattern of hybridization; furthermore, weak hybridization was detectable between the two different classes. Inserts from one clone of each H9/HTLV-III-specific class (H9c.7 and H9c.53) were subcloned into phage M13 vectors and their sequences determined by the dideoxy-chain terminator method. Significantly, a 76-nucleotide sequence was shared by the 5′ end of H9c.7 and the 3′ end of H9c.53, accounting for hybridization between the two classes. The 3′ point of divergence of this 76-nucleotide sequence was marked by a polyadenylate tract in H9c.53, as expected for an RNA polymerase II transcript. The congruence exhibited by the opposite ends of these clones was very like the terminal redundancy of the viral genome of retroviruses[22-24], suggesting that clones H9c.7 and H9c.53 represented the 5′ and 3′ regions, respectively, of HTLV-III (Fig. 1).

The identity of H9c.7 and H9c.53 was confirmed by blot hybridization analysis of H9/HTLV-III and normal human lymphocyte genomic DNA restriction digests. Sequences hybridizing with H9c.7 and H9c.53 were found only in DNA from infected cells, demonstrating their exogenous viral origin (Fig. 2*A*). To determine whether related sequences are associated

* Present address: Invitron, 8000 Maryland Avenue, Clayton, Missouri 63105, USA.

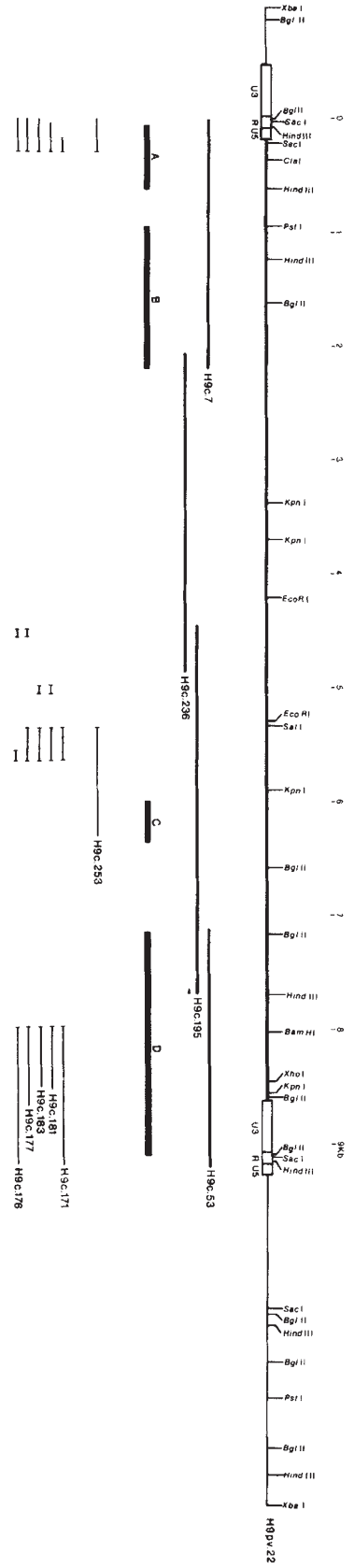


Fig. 1 Map of LAV/HTLV-III proviral and cDNA clones. Proviral sequences (bold line), flanking cellular DNA (thin line), LTR regions (boxed areas) and restriction endonuclease sites are indicated at top. Sequences are numbered from the start of viral transcription. cDNAs representing the full-length viral RNA (shown immediately below the provirus) and spliced subgenomic RNAs (at bottom of figure) are shown with the respective clone designation to the right of its map (H9pv, proviral λ clone; H9c, viral cDNA clone). The horizontal bars labelled A, B, C and D refer to fragments used as probes for the identification of spliced cDNA clones. The arrowhead at nucleotides 7,687–7,702 denotes the 16-mer primer, 5′-CTCTGTCCCACTCCAT, used to prime cDNA synthesis of clone H9c.195.

Methods. Viral cDNA clones were isolated from a cDNA library prepared from H9/HTLV-III poly(A)$^+$ RNA. Double-stranded cDNA was synthesized as described elsewhere[52] and synthetic *Eco*RI adaptors were added[53]. After the removal of excess adaptor, insertion into λgt10[54] and *in vitro* packaging, the recombinant phage were plated at a density of $\sim 5 \times 10^3$ plaques per 150-mm plate. Replica nitrocellulose filters were made from the plates[55] and the library probed with the product of uniformly ^{32}P-labelled first-strand cDNA prepared as described above from poly(A)$^+$ RNA from either H9/HTLV-III cells or the uninfected human T-cell line CEM[56]. The filters were hybridized at 42 °C in a solution containing 5×SSC, 50 mM sodium phosphate (*p*H 6.8), 0.1% sodium pyrophosphate, 5× Denhardt's solution, 0.04 g l^{-1} sonicated salmon sperm DNA, 50% formamide and 10% dextran sulphate and washed at the same temperature in 0.2×SSC, 0.1% SDS. Plaques hybridizing specifically with the H9/HTLV-III probe were plaque purified and their DNA prepared for further analysis. To isolate clones comprising the remainder of the viral RNA genome, a synthetic oligodeoxynucleotide positioned near the 5′ end of clone H9c.53 was used to prime specifically cDNA synthesis. One clone obtained from the resulting cDNA library, H9c.195, extended for 3.2 kb beyond the primer location but did not overlap into the region represented by H9c.7. Rescreening of the library with the H9c.195 insert allowed the recovery of a 3.5-kb clone, H9c.236, which covered the remaining viral sequences. To isolate clones containing the complete integrated provirus, *Xba*I-digested H9/HTLV-III genomic DNA was fractionated on sucrose gradients and the resulting 10–20-kb fragments were inserted into the *Xba*I cloning site of the bacteriophage λ vector J1 (ref. 57). Twenty-six clones containing an integrated provirus were recovered from $\sim 1 \times 10^6$ recombinants screened with clones H9c.7 and H9c.53.

with LAV infection, total DNA isolated from peripheral blood lymphocytes acutely infected with LAV was hybridized with H9c.7 and H9c.53 under stringent conditions. Fragments of similar sizes were detected in *Hin*dIII, *Bgl*II and *Sst*I digests of DNA from H9/HTLV-III and LAV-infected cells (Fig. 2*A*). This result demonstrates clearly that HTLV-III and LAV correspond to the same or very closely related viruses. By contrast, H9c.7 and H9c.53 did not hybridize to cloned HTLV-I or HTLV-II provirus sequences, even in conditions of low stringency (data not shown). Significantly, the level of hybridization detected with DNA from LAV-infected lymphocytes was at least 20-fold greater than that with H9/HTLV-III DNA (Fig. 2*A*). Most of the hybridization observed with LAV-infected cell DNA migrates as a 9.5-kilobase (kb) species in *Xba*I digests (Fig. 2*A*, lane *h*) or with undigested DNA (data not shown), suggesting that this represented linear unintegrated viral DNA. By contrast, little unintegrated DNA was detected with H9/HTLV-III; instead, most of the DNA in the *Xba*I digest appeared in five or more discrete species of integrated proviruses 15–20 kb in size (Fig. 2*A*, lane *g*). Dot-blot hybridizations confirmed the presence of 5–10 copies of proviral DNA in H9/HTLV-III cells (data not shown). As no more than 5% of cells treated with LAV seemed to be infected by the virus by indirect immunofluorescence of viral antigens (data not shown), there thus seemed to be several hundred copies of unintegrated DNA present per LAV-infected cell.

Clones comprising the remainder of the viral RNA genome (H9c.236, H9c.195) were identified in a second cDNA library prepared with a specific primer (see Fig. 1). The regions represented by these four overlapping cDNA clones and their restriction maps are shown in Fig. 1. The size predicted for the full-length viral RNA genome, 9.2 kb, is consistent with the largest species observed by blot analysis of H9/HTLV-III poly(A)$^+$RNA (Fig. 2*B*). Given the possibility that the cDNA clones isolated might reflect RNA splicing events leading to the removal of small introns and therefore do not represent the entire viral genome, it was necessary to isolate molecular clones

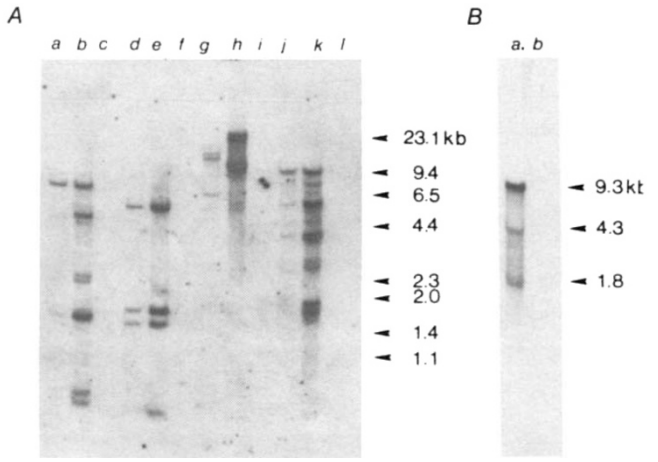


Fig. 2 Blot hybridization analysis of DNA and RNA from HTLV-III- and LAV-infected cells. *A*, Genomic DNA was isolated from the H9/HTLV-III cell line, acutely LAV-infected peripheral lymphocytes or normal blood leukocytes. Each (5 μg) was digested with the indicated restriction endonuclease, electrophoresed on a 1% agarose gel and transferred to nitrocellulose[58]. The blot was hybridized with ^{32}P-labelled probes[59] A, B and D (see Fig. 1) of equal specific activity using the hybridization and washing conditions described in Fig. 1 legend. The fragments obtained from *Hin*dIII digestion of λDNA and *Hae*III digestion of ΦX174 were used as relative molecular mass markers; these sizes are shown in kilobase pairs (kb) on the right. Lanes *a, d, g, j*, HTLV-III-infected H9 DNA; lanes *b, e, h, k*, LAV-infected lymphocyte DNA; lanes *c, f, i, l*, normal leukocyte DNA. DNA in lanes *a–c* was digested with *Hin*dIII: *d–f*, *Bgl*II; *g–i*, *Xba*I; *j–l*, *Sst*I. ***B***, 1 μg poly(A)$^+$ RNA from either HTLV-III-infected H9 cells (*a*), or non-infected H9 cells (*b*) was electrophoresed on a 1% formaldehyde/agarose gel[60], transferred to nitrocellulose and hybridized to ^{32}P-labelled probe D (see Fig. 1) in hybridization and washing conditions identical to those in Fig. 1 legend. The size of each predominant RNA species was estimated from migration of single-stranded DNA markers and is indicated on the right.

of the integrated provirus. A library of size-enriched DNA was constructed to isolate proviral clones taking advantage of absence of *Xba*I sites in the provirus apparent from both the Southern blot results (Fig. 2*A*) and the map of the cDNA clones (Fig. 1). The restriction map of the integrated provirus clone selected for sequence analysis, H9pv.22, showing the proviral and adjoining cellular sequences is presented in Fig. 1.

Nucleotide sequence analysis

The complete DNA sequence of the integrated provirus is presented in Fig. 3 and compared with that determined for the four overlapping cDNA clones. The co-linearity of these sequences confirms that the cDNA clones isolated represent the unspliced viral genomic RNA. A significant degree of nucleotide heterogeneity is displayed by the proviral and cDNA sequences. Similarly, nucleotide differences are evident in the overlap regions between cDNA clones. This genomic diversity probably reflects the observation that the H9/HTLV-III cell line studied here was established by infection with material from several AIDS patients[10]. Together with the blot hybridization data indicating the presence of about five intact provirus copies (Fig. 2*A*, lane *g*), these results suggest that two or more distinct virus isolates are integrated stably in the H9/HTLV-III cell line. Despite 68 nucleotide differences between the provirus and cDNA sequences, a consistent structure nevertheless emerges for the coding potential of the virus.

The long terminal repeats (LTRs) of the provirus exhibit the structural features common to all retroviral LTRs which reflect the manner of viral replication and are essential to the mode of entry of the virus into the host genome and recognition by the cellular transcriptional apparatus[25]. The unique 5′ (U5) and 3′ (U3) regions are bordered by sequences complementary to the

Fig. 3 (Right) Complete nucleotide sequence of the HTLV-III/LAV provirus genome. The sequence shown represents the coding strand, comprising 9,213 bp extending from the RNA cap site to the polyadenylation site. The following features of the DNA sequence are indicated: the U5, R and U3 regions; splice acceptors (^{sa}) and splice donors (^{sd}); the inverted repeat located at the 5′ end of U3 and the 3′ end of U5 (IR); the tRNA^{Lys} primer binding site (PBS) and the (+)-strand initiation site (+). The polyadenylation signal (AATAAA) and Goldberg-Hogness sequence (TATAAG) are boxed. Deduced amino acid sequences of *gag, pol, P′, env*, and *E′* are shown above the DNA. The NH_2-terminus of p24^{gag} and that predicted for gp65^{env} and gp41^{env} (see text) are indicated. Nucleotide variations between cDNA and proviral isolates are below the line, and the resulting amino acid differences are above. The three nucleotides (TAA) indicated below the line at position 56–62 represent an insertion in the cDNA sequence.
Methods. cDNA inserts were mapped with restriction endonucleases, fragments isolated and cloned into M13 vectors[61]. Single-stranded template was isolated and the sequence determined using the chain termination method[62]. Additional fragments were sequenced to determine the overlap junctions. Using the completed cDNA sequence, overlapping fragments of ~800–1,000 nucleotides were isolated from the provirus clone H9pv.22 for comparative sequence analysis.

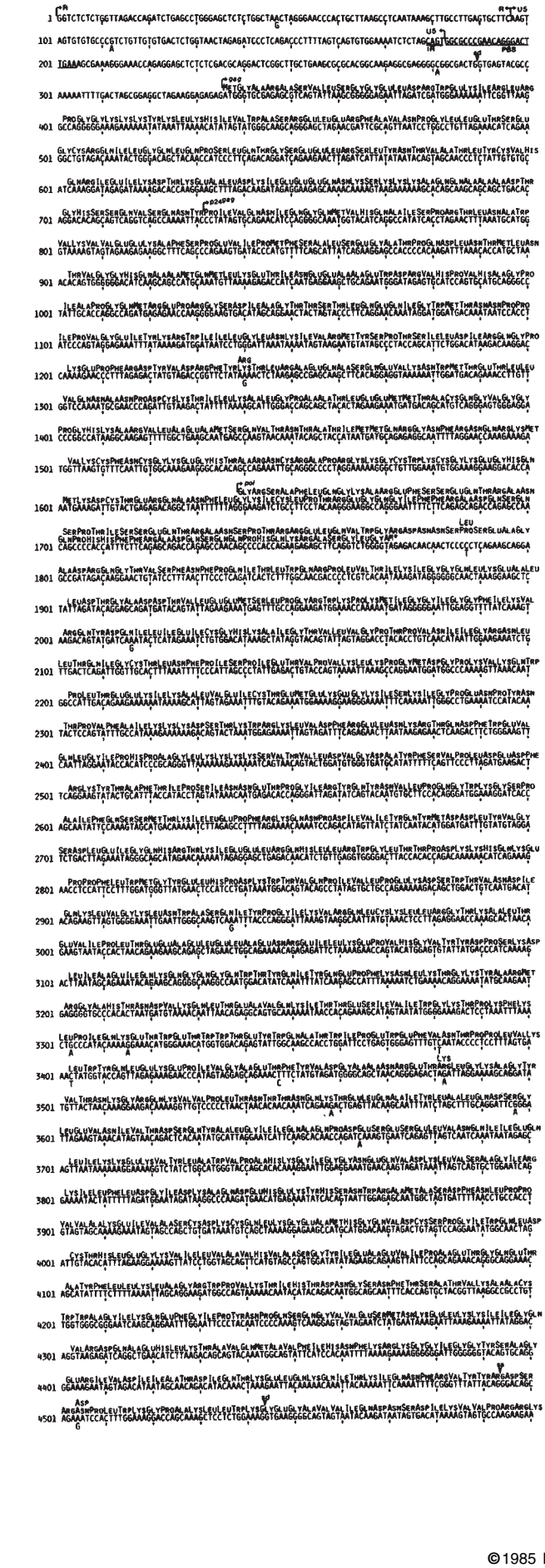
primers which copy the terminally redundant sequences (R) of the viral RNA to accomplish the series of intermolecular strand exchanges responsible for viral (−)- and (+)-strand DNA synthesis. These borders correspond to the ends of the resulting linear duplex molecule from which, in all cases so far examined, two base pairs (bp) are lost on insertion of the provirus into cellular DNA[25]. Based on this premise, the 5′ end of U3 (nucleotide 8,662) and the 3′ end of U5 (nucleotide 182) were identified from the sequence of the virus-cell DNA junctions in the H9pv.22 provirus clone (see Fig. 1). As predicted, a 23-bp sequence 3′ to the U5 boundary is complementary to a potential primer for (−)-strand synthesis, transfer RNA^{Lys} the same primer used by the mouse mammary tumour virus (MMTV)[26], whereas that 5′ to the boundary U3 consists of a stretch of 15 purines, the sequence found generally at the site of (+)-strand chain initiation[25].

In accord with the general retroviral paradigm, the integration event represented by the H9pv.22 clone reflects a duplication of host sequences at the insertion site; a short inverted repeat is found at the ends of the viral LTR. The virus-cell DNA junction sequence, TGTAGTGGGTG...CAGTGGGTGAT (viral sequences underlined), indicates a 5-bp duplication (GTGGG) of cellular DNA and an inverted repeat of 4 bp (ACTG...CAGT), two nucleotides of which are lost on insertion. In agreement with the general finding that the size of the duplication resulting from integration is a property of the virus and not the host cell, it is interesting that integration of HTLV-I, which shares OKT4 tropism with LAV/HTLV-III, results in a direct repeat of 6 bp of cellular DNA[27].

The 3′ end of the viral RNA (Fig. 3) corresponds to the site of poly(A) addition in the H9c.53 cDNA clone and is 18 nucleotides from the polyadenylation signal, AATAAA[28]. The 5′ end of the viral RNA was mapped by *in vitro* extension of a synthetic DNA primer complementary to sequences in U5. As shown in Fig. 4, most of the RNA is initiated at the G residue indicated as position 1 (Fig. 3), although heterogeneity of two or three nucleotides is observed. Based on these results, the sizes of the U3, R and U5 regions are 456, 96 and 86 nucleotides, respectively. The RNA initiation site is located 23 nucleotides from the sequence TATAAG, which conforms to the consensus Goldberg-Hogness box found characteristically in this distance and implicated in the positioning of eukaryotic transcription initiation sites[29]. A direct repeat of 10 bp is found 54 nucleotides 5′ to this sequence (nucleotides 9,013–9,022; 9,027–9,036).

The *gag* gene encodes p24

The *gag* reading frame indicated by the provirus sequence extends from nucleotides 336–1,769 and, by analogy to other retroviruses, is expected to code for a precursor polypeptide



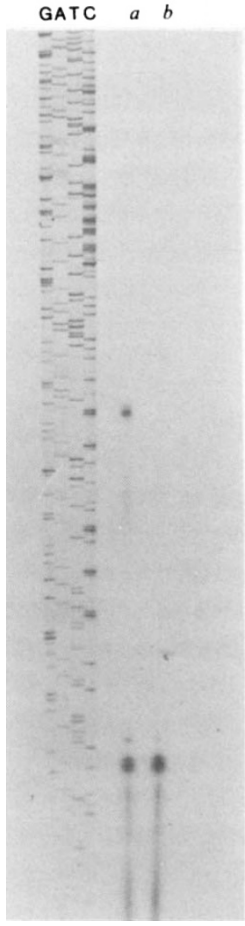


Fig. 4 Determination of the initiation site of viral transcription. Autoradiograph of a 5% polyacrylamide, 8 M urea gel. GATC, a DNA sequence ladder corresponding to the region surrounding the U3/R junction of the 5′ LTR. Results of primer extension reactions are shown using as template poly(A)$^+$ RNA prepared from HTLV-III-infected (*a*) or uninfected (*b*) H9 cells.

Methods. A synthetic 16-mer complementary to the sense (+)-strand in U5 (position 129–144, see Fig. 3) was annealed to a phage M13 single-stranded template containing a portion of the (+)-strand corresponding to U5, R and part of U3. The primer was extended by incubation at 37 °C for 30 min in the presence of [α^{32}P]dCTP, [α^{32}P]dATP, dTTP, dGTP and DNA polymerase I (Klenow fragment). The product was digested with *Hin*dIII, heated at 100 °C for 5 min and fractionated on a 6% polyacrylamide gel. The labelled, 63-base-long, single-stranded fragment extending from position 144 to the *Hin*dIII cleavage site at position 81 was recovered; ~5×10^5 c.p.m. ^{32}P was annealed to 7.5 μg poly(A)$^+$ RNA in 80% formamide, 0.4 M NaCl, 40 mM PIPES (*p*H 6.5), 1 mM EDTA at 52 °C for 3 h. After primer extension with avian myeloblastosis virus reverse transcriptase, the sample was treated with 100 mM NaOH at 70 °C for 45 min and loaded onto a 5% polyacrylamide, 8 M urea gel. The DNA sequence used as marker was generated by using the above 16-mer as the primer in dideoxy method sequencing reactions on the same M13 template used to prepare the 63-bp fragment, therefore the primer extension products co-migrate with the corresponding nucleotide sequence.

which is post-translationally cleaved to give the internal structural proteins of the virus. Translation of this reading frame begins at the 5′-proximal ATG triplet of the viral genomic RNA and would lead to the synthesis of a 478 amino acid polypeptide consistent with the relative molecular mass (M_r) 53–55,000 of the *gag*-encoded polyprotein (Pr54^{gag}) detected recently in H9/HTLV-III cells[30]. A good candidate for the major virion core protein encoded by this precursor is a 24,000 M_r virus-associated protein (p24) recognized by sera from AIDS and ARC patients[7–9,15,16] (see Fig. 6). Recently, the N-terminal amino acid sequence of p24 has been determined (J. Bell and C.V.B., unpublished observations) and the 17-residue sequence obtained for the purified protein matches exactly that beginning with the proline residue found at position 732 (Fig. 3).

The p24^{gag} N-terminal cleavage thus identified predicts a 132 amino acid protein from the N-terminus of the *gag*-encoded polyprotein, containing a single potential asparagine-linked *N*-glycosylation site. Interestingly, this cleavage recognition site (the aromatic amino acid proline) is analogous to that found in the precursor for Moloney murine leukaemia virus (Mo-MuLV)[31], suggesting that processing occurs by a viral or cellular protease with similar specificity.

Although the proteolytic cleavage responsible for generating the C-terminus of p24^{gag} has not yet been defined, the presence of ~130 amino acids beyond the sequence sufficient to encode p24^{gag} indicates that a third protein is encoded by Pr54^{gag}. A direct repeat of 36 nucleotides resulting in a C-terminal duplication of 11 amino acids (positions 1,676–1,747) is obvious in this region. The sequence of this protein shows significant conservation of cysteine residues with p12^{gag} of Rous Sarcoma Virus (RSV), p11^{gag} of HTLV-I and p10^{gag} of Mo-MuLV[27,31,32]. Common to each protein are three cysteine residues separated by two and nine amino acids, respectively; this structure is found twice in the HTLV-III, RSV and HTLV-I proteins and once in Mo-MuLV. The p12^{gag} of RSV, a protein rich in basic residues, is the major component of the virion ribonucleoprotein complex, binding nonspecifically to many sites on the viral RNA[33,34]. Similarly, the HTLV-III protein is highly basic, containing 23 lysine and arginine residues. These striking similarities between otherwise highly-diverged proteins suggests that this C-terminal *gag*-encoded protein constitutes the core ribonucleoprotein.

The *pol* region

The product of the *pol* gene is encoded by the second large reading frame of the virus located between nucleotides 1,639–4,674. The predicted amino acid sequence shows significant homology to the *pol* gene products of RSV, HTLV-I, and Mo-MuLV (~20–30%)[27,31,32], confirming the assignment of this reading frame to the *pol* gene. Despite the conservation of amino acid structure, however, there is no significant corresponding nucleotide homology with RSV, HTLV-I or Mo-MuLV, in contrast to earlier reports demonstrating cross-hybridization between HTLV-III and HTLV-I in this region[35,36]. The reverse transcriptase of several avian and mammalian retroviruses is translated initially as a *gag–pol*-encoded polyprotein, an event for which either the suppression of an in-frame amber codon (Mo-MuLV) or the removal of a short interval by RNA splicing (RSV) has been suggested as a mechanism for joining the reading frames of the *gag* and *pol* genes[31,32]. The *gag* and *pol* genes of HTLV-III are also situated in different reading frames (Fig. 3), suggesting a similar requirement for splicing as a mechanism of *pol* expression. The first ATG codon encountered in the *pol* gene is at position 1,939; however, the potential reading frame begins at position 1,639, overlapping the *gag* gene by 130 nucleotides. Thus, the possibility exists for shared amino acid sequences between the C-terminus of the *gag*-encoded ribonucleoprotein and the N-terminus of reverse transcriptase.

There exists a third reading frame, designated *P′*, between nucleotides 4,589–5,197, containing ATG triplets at positions 4,622, 4,643, 4,667 and 4,706 and could thus encode a protein of ~192 amino acids. The potential for generating spliced transcripts containing *P′* sequences suggests that this reading frame is used (see below). Beyond *P′* there is an intercistronic region of 602 nucleotides with frequent stop codons in all three reading frames. This contrasts with the compact organization of Mo-MuLV and HTLV-I that have overlapping *pol* and *env* genes[27,31].

The *env* gene

The primary translation product of the *env* gene of retroviruses is a large glycosylated precursor synthesized on the rough endoplasmic reticulum, which is processed to produce a larger N-terminal glycoprotein bearing the host range determinants and a smaller hydrophobic protein serving as the membrane anchor by virtue of a transmembrane domain. The *env* gene product

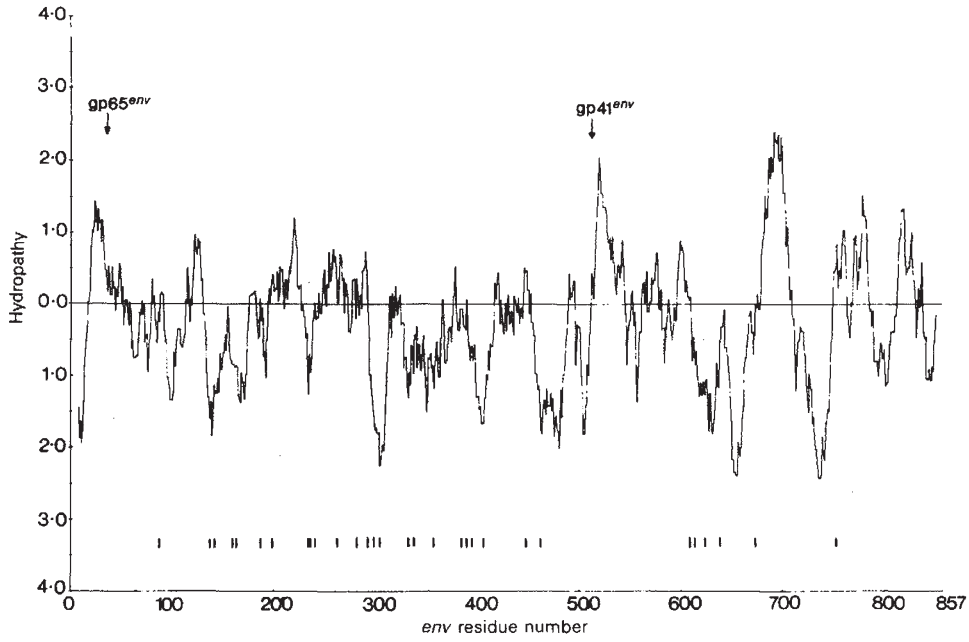


Fig. 5 Hydropathy plot for the *env* gene product. Hydrophobic areas appear above the midline, and hydrophilic areas below. Potential sites of asparagine-linked glycosylation are indicated by vertical lines below the plot. The NH_2-termini of the putative mature $gp65^{env}$ and $gp41^{env}$ are indicated. The region corresponding to $gp65^{env}$ spans ~480 amino acids beginning after a hydrophobic leader and extending to the conserved processing site Arg-N-(Arg/Lys)-Arg which marks the NH_2-terminus of $gp41^{env}$. The $gp41^{env}$ region consists of 345 amino acids containing two extended hydrophobic domains. Hydropathy was calculated by the method of Kyte and Doolittle[63] for each overlapping segment of 15 amino acids.

of HTLV-III is encoded by the large open reading frame at nucleotides 5,782–8,370. Analysis of a cDNA clone representing the *env* mRNA (see below) suggests that translation of the *env* region is initiated at the ATG codon located near the beginning of this reading frame at position 5,803. This assignment predicts the synthesis of an 856-residue envelope precursor protein containing 30 potential sites of asparagine glycosylation located principally in the first half of the molecule (Fig. 5). Taking into account the addition of carbohydrate, the predicted size of the *env* precursor approximates gp120, a 120,000 M_r glycoprotein detected recently in intracellularly labelled H9/HTLV-III cells[30].

Although N-terminal sequence information is presently unavailable for the envelope proteins of HTLV-III and there is no discernible amino acid homology with the *env* gene products of RSV, Mo-MuLV, HTLV-I or MMTV[27,31,32,37], certain features of the sequence allow the prediction of the processing of the precursor molecule. Three stretches of hydrophobic and nonpolar residues are present at positions 5,851–5,886 (12 residues), 7,336–7,419 (28 residues) and 7,852–7,917 (22 residues) (Fig. 5). The first stretch is located at the N-terminus of the precursor and is flanked by charged residues, suggesting a role as the signal sequence responsible for directing the protein to the cell surface[38]. The size and position of the other two hydrophobic regions suggest that they are located in the transmembrane protein of the HTLV-III envelope. The transmembrane envelope proteins of RSV, Mo-MuLV, MMTV and HTLV-I similarly display two hydrophobic stretches of 20–30 amino acids separated by ~150–200 residues[27,31,32,37]. In each case, maturation of the envelope polyprotein involves cleavage after a conserved sequence of basic residues, Arg-N-(Arg/Lys)-Arg, immediately preceding the first hydrophobic stretch. The presence of this sequence at the corresponding position of the HTLV-III *env* gene product suggests that $gp120^{env}$ is cleaved into a N-terminal protein of ~480 amino acids and a transmembrane protein of 345 amino acids with 24 and 6 potential asparagine-linked glycosylation sites, respectively (Fig. 5).

This assignment for the major envelope glycoprotein and transmembrane protein is supported by serological evidence. We analysed HTLV-III virion proteins by immunoblotting with antisera from individuals infected with the virus. With extensively diluted sera from infected individuals, we detect a predominant specis of M_r 65,000 in addition to $p24^{gag}$ (Fig. 6*A–E*), which does not react with undiluted sera from normal individuals (Fig. 6*F*) suggesting that p65, like $p24^{gag}$, is a major structural constituent of the virion. The size of p65 is consistent with the predicted size of the major envelope glycoprotein. Previous studies have indicated the presence of a p65 in HTLV-III virion preparations and in H9/HTLV-III cells[15,16,39], but have consistently detected far greater amounts of a 41,000 M_r glycoprotein[16,17] (also detected by the sera analysed in Fig. 6). In light of the present evidence, it seems that gp41 may represent the transmembrane protein rather than the major glycoprotein. The apparent difference in the ability of patient sera to detect p65 rather than gp41 in these studies may reflect the method of virus preparation.

Beyond the *env* gene there is a fifth open reading frame, designated *E'*, between nucleotides 8,347–8,992, extending into the U3 region of LTR. This novel reading frame can encode a protein of 206 amino acid residues, beginning at the ATG triplet at nucleotide 8,370, which is unrelated to the X genes of HTLV-I or HTLV-II (refs 27, 40, 41).

Transcription analysis

The synthesis of proteins encoded at internal sites in the retroviral genome is accomplished by two mechanisms; the post-translational cleavage of polyprotein precursors and for genes located in the 3′ half of the genome (the *env* gene and transforming genes of several acute transforming viruses), by the expression of spliced subgenomic messenger RNAs. Northern analysis of H9/HTLV-III RNA revealed three prevalent viral specific RNAs of 9.3, 4.3 and 1.8 kb (Fig. 2*B*), suggesting the expression of three major virus-encoded primary translation products. To determine whether the 4.3- and 1.8-kb subgenomic RNAs could account for the synthesis of the HTLV-III *env* and *E'* gene products, viral transcription was analysed with hybridization probes specific for the detection of cDNA clones representing spliced transcripts. Evidence from several retroviruses indicates that sequences from the 5′ end of the viral genome are found in each of the major viral RNAs[42–44]. Therefore, we sought cDNA clones that would hybridize to probes derived from the 5′ end of the virus (probe A) as well as the *env* (probes C, D) or *E'* regions (probe D), but not with sequences corresponding to the *gag–pol* region (probe B) (Fig. 1).

This screening strategy allowed the identification of two distinct classes of subgenomic mRNA clones. The structures of these transcripts (obtained by DNA sequence analysis) are presented in Fig. 1. The first class of spliced mRNAs, typified by clone H9c.253, is comparable to the RNA species encoding the envelope proteins of other retroviruses[45]. Transcription of this class of mRNA is initiated in the 5′ LTR, probably extends to the polyadenylation site in the 3′ LTR and reflects the removal of a large intron between nucleotides 289–5,359 containing the

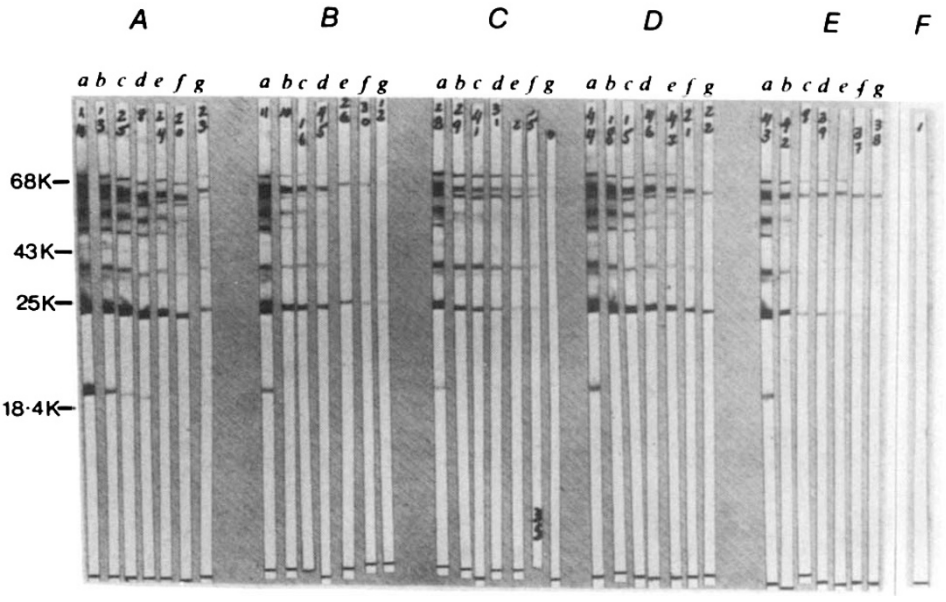


Fig. 6 Immunoblot detection of viral antigens. Virion-specific proteins were detected by sera from two healthy homosexual men (*A, D*), plasma from two other healthy homosexual men (*B, C*), serum from a sexual partner of an AIDS patient (*E*), but not by undiluted serum from a healthy control individual (*F*). Serum samples were undiluted (*a*), diluted 1:10 (*b*), 1:20 (*c*), 1:40 (*d*), 1:80 (*e*), 1:160 (*f*), 1:320 (*g*). At the highest sample dilutions, only p24 and p65 are detectable.
Methods. HTLV-III-infected H9 cells were grown in RPMI 1640 supplemented with 5% fetal bovine serum. Virus particles were concentrated from culture fluids by ultrafiltration and purified by banding on a 20–60% sucrose gradient in an SW 27 rotor centrifuged for 16 h at 27,000 r.p.m. Fractions between 30–40% were diluted and the viral particles pelleted by ultracentrifugation. Viral proteins were separated on a 12% SDS-polyacrylamide gel and transferred to nitrocellulose. The blot was cut into strips and incubated overnight at 25 °C with the patient sample. After washing, virus-specific protein bands were visualized by incubation with biotinylated goat anti-human IgG and subsequent incubation with avidin–horseradish peroxidase. M_r standards of 68, 43, 25 and 18.4 ($\times 10^3$) are shown.

gag, *pol* and *P'* genes. The size predicted for this RNA, 4,144 nucleotides plus poly(A), corresponds to the 4.3-kb species detected by Northern analysis (Fig. 2*B*). The putative *env* initiation codon is preceded by three ATG triplets at positions 5,412, 5,551 and 5,643; however, each is followed by a nearby in-frame stop codon located 216, 81 and 243 nucleotides 3', respectively. The selection of an ATG codon other than the most 5' proximal ATG triplet for translation initiation has been previously noted for *gag* and *env* expression in Mo-MuLV and RSV[31,32].

The second class of spliced mRNAs identified hybridized with probe D but not probe C, indicating that they did not contain most of the *env* gene. This class of mRNAs reflected a variety of splicing events, involving the removal of two or more introns (Fig. 1) with the generation of an RNA species similar in size (~1.8 kb) to the smaller subgenomic RNA seen by blot analysis (Fig. 2*B*). Each member of this class was formed from the 5' leader (exon 1) and sequences corresponding to the first 268 nucleotides of *env* leader (exon 4). In every case, the latter sequences were joined directly to the splice acceptor at position 7,957 in the C-terminal coding region of *env* (exon 7). In addition, several clones possessed an additional untranslated leader exon of 50 nucleotides (exon 2) (H9c.181, H9c.183) or 74 nucleotides (exon 3) (H9c.176, H9c.177) derived from the *pol* and *P'* regions, respectively. Table 1 summarizes the size and location of each exon and compares their donor and acceptor sequences. As a consequence of the splicing event which deletes most of the *env* gene, the ATG triplet at position 5,643 is removed whereas the distance between the ATG triplets at positions 5,412 and 5,551 and the next in-frame stop codons increases to 258 and 348 nucleotides, respectively. Neither of these reading frames connects with the sequences that encode the C-terminus of the *env* polyprotein. One of the clones (H9c.176) uses an alternative splice acceptor in exon 4, resulting in the juxtaposition of exon 2 with the last 69 nucleotides of the *env* leader (exon 5) and the removal of the ATG codons at 5,412 and 5,551. Significantly, the members of this diverse class of spliced RNAs are all related by their ability to direct the synthesis of the putative *E'* gene product.

Discussion

Although a definitive interpretation of the HTLV-III sequence awaits the demonstration of biological activity for our virus clones, the concordance of provirus and cDNA sequences corresponding to apparently distinct virus isolates suggests that they portray accurately the structural features of the viral genome. Despite a gene organization in many aspects similar to other retroviruses, the HTLV-III genome seems to be entirely unrelated by nucleotide homology to previously characterized retroviral sequences[27,31,32,37], including HTLV-I and HTLV-II which display a similar tropism for the OKT4$^+$ T-cell subset[46]. The

Table 1 Summary of LAV/HTLV-III exon and splice junctions

Exon	Location	Length (nucleotides)	Splice acceptor	Splice donor
1	1–289	289	– ...	GACTG|GTGAGTAC
2	4,494–4,543	50	TTTCGGGTTTATTACAG|GGA ...	GAAAG|GTGAAGGG
3	4,971–5,044	74	CTTTGACTGTTTTTCAG|ACT ...	ACAAG|GTAGGATC
4	5,359–5,626	268	TGTTTATCCATTTTCAG|AAT ...	AAGCA|GTAAGTAG
5	5,558–5,626	69	GGCATCTCCTATGGCAG|GAA ...	AAGCA|GTAAGTAG
6	5,359–9,213	3,855	TGTTTATCCATTTTCAG|AAT ...	–
7	7,957–9,213	1,257	CACCATTATCGTTTCAG|ACC ...	–

Each row lists the known exons (see Fig. 7), the nucleotide positions comprising each exon (see Fig. 3), and total exon length. The DNA sequence adjacent to the 5' (acceptor) and 3' (donor) borders at each exon are also shown. Intron sequences are to the left of the vertical line in the acceptor column, and to the right of the vertical line in the donor column. These sequences were obtained by comparison of the complete proviral sequence with those of several cloned cDNAs representing spliced subgenomic mRNAs, shown in Fig. 1.

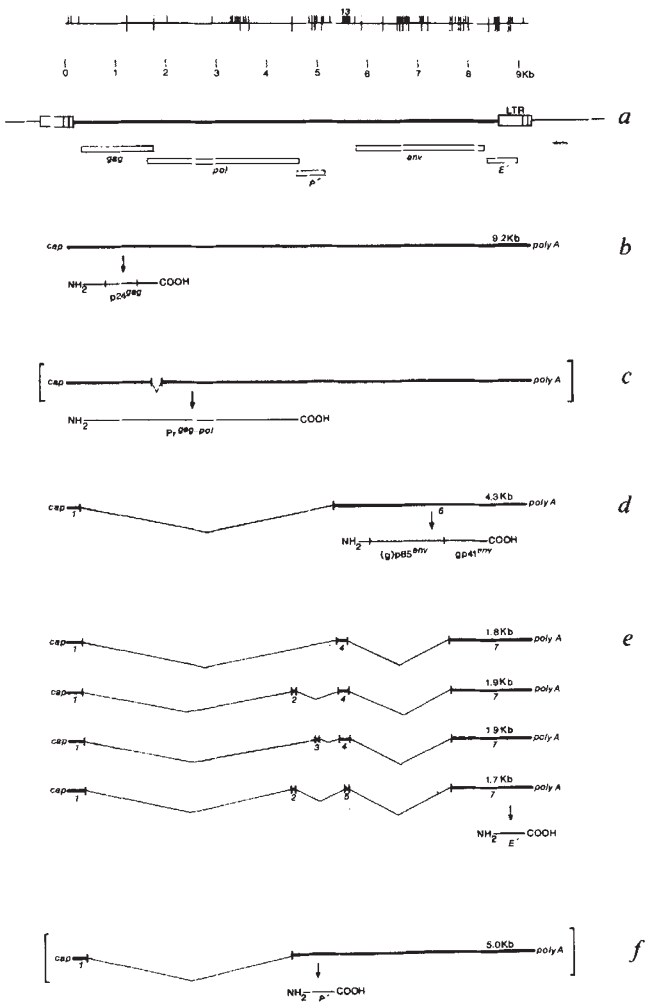


Fig. 7 Summary of viral gene organization and expression. *a*, Translational reading frames for the viral genes are indicated below a representation of the integrated proviral genome. Vertical marks above the line denote the location of nucleotide differences between the proviral (H9pv.22) and cDNA clones (H9c.7, H9c.236, H9c.195, H9c.53; vertical marks below the line indicate the resulting amino acid substitutions. *b*, Synthesis and subsequent processing of the *gag* precursor from full-length viral RNA. *c*, Hypothetical splicing pathway similar to that proposed for RSV[32] which would allow for synthesis of a *gag-pol* fusion protein. *d*, Synthesis and processing of the *env* polyprotein from the 4.3-kb spliced, subgenomic mRNA. *e*, Predicted synthesis of *E'* gene product from the 1.8-kb class of spliced, subgenomic mRNAs. *f*, Hypothetical subgenomic RNA joining the 5' leader with exon 2 without subsequent splicing that could encode the predicted *P'* gene product.

most significant protein homology detected between the gene products of HTLV-III and other retroviruses is in the *pol* gene (~20–30%); however, a corresponding degree of nucleotide relatedness is not found, which is inconsistent with reports of homology to HTLV-I in the *gag-pol* region based on hybridization data[35,36]. These results and the observation that the virus is morphologically distinct from the type C viruses[8,13,15] lead us to propose that LAV/HTLV-III is a member of a novel class of retroviruses, perhaps including the equine infectious anaemia virus, which has a similar morphology and a serologically-related major core protein[8].

Figure 7 summarizes the gene organization, transcriptional processing and polyprotein maturation pathways predicted for the virus. The *gag* gene product is a precursor of M_r 54,000 which, as found for other retroviruses, seems to be synthesized from genome-size viral mRNA (Fig. 7*b*). $Pr54^{gag}$ is cleaved at two positions to generate the major core structural protein, $p24^{gag}$, a M_r ~ 15,000 N-terminal polypeptide and a structurally-conserved C-terminal ribonucleoprotein of M_r ~ 15,000. The *pol* and *gag* genes are encoded by different reading frames; translation of a *gag-pol*-encoded polyprotein would thus require joining of the reading frames by a splicing event (Fig. 7*c*).

Characterization of a cDNA clone which probably represents the 4.3-kb major viral subgenomic RNA has allowed the identification of the transcript capable of directing the stoichiometric synthesis of the *env* gene products. The removal of a single intron containing the *gag-pol* region accomplishes the joining of a 289-nucleotide leader from the 5' end of the genome with a second untranslated leader situated within an intercistronic region between *pol* and *env* (Fig. 7*d*). Translation of this spliced subgenomic RNA would lead to synthesis of an 856 amino acid envelope polyprotein with a hydrophobic signal leader which would direct the precursor to the cell membrane and allow initiation of the viral envelope formation. Evidence of a conserved cleavage recognition site preceding an extended hydrophobic domain suggests processing of $Pr120^{env}$ generates the major envelope and transmembrane glycoproteins. The sizes thus predicted for $gp65^{env}$ and $gp41^{env}$ are ~480 and 345 residues, respectively, in agreement with the observed electrophoretic mobilities of two major virion structural proteins. Demonstration of the glycoprotein nature of p65 or direct sequence analysis of these polypeptides will be required to affirm these assignments.

Expression of the novel *E'* gene product is indicated by a class of cDNA clones corresponding to several related but structurally heterogeneous spliced subgenomic RNAs of ~1.8 kb (Fig. 7*e*). The salient feature of this class is a tripartite structure consisting of the 5' leader, the intercistronic leader and 1.3 kb of RNA from the 3' end of the genome. Translation of the *E'* reading frame would result in the synthesis of a 206-residue protein lacking homology with the *X* gene products encoded by the 3' regions of HTLV-I and HTLV-II. Differentially-spliced mRNAs capable of encoding the *E'* gene product are generated by the addition of either of two additional untranslated leaders or by the use of an alternative splice acceptor in the intercistronic leader sequence. Although the existence of *E'* is unprecedented in other retroviral genomes, it is interesting to speculate that it acts as a virus-specific transcription factor, a function ascribed recently to the HTLV-I and HTLV-II *X* gene products[47,48]. The multiplicity of splicing patterns also provides a rationale for the generation of a subgenomic mRNA directing the synthesis of the gene product encoded by the *P'* reading frame (Fig. 7*f*). In the examples already described, the joining of the 5' leader to the small exon in the *pol* gene (exon 2) is accompanied always by further splicing at the donor site located 50 nucleotides 3'. If further splicing at this donor did not occur, a subgenomic RNA ~870 nucleotides longer than the *env* mRNA would be produced having as its primary translation product a 192 amino acid protein specified by the *P'* reading frame. Alternatively, the removal of a small intron between the *pol* and *P'* reading frames could result in the synthesis of a *gag-pol-P'* precursor peptide. Further investigation will be required to establish the identity the *E'* and *P'* gene products and their roles in viral reproduction and pathogenesis.

Following retroviral infection, the major product of viral DNA synthesis in the cytoplasm is a linear, double-stranded molecule of genome size with a complete copy of the LTR at each end[25]. Studies with several cytopathic retroviruses, notably spleen necrosis virus and the cytopathic subgroups of avian leukosis virus, have revealed a correlation between transient cell killing during acute infection and the transient accumulation of 100–200 copies of linear unintegrated viral DNA[49–51]. Both transient cell killing and transient accumulation of linear unintegrated DNA require the spread of virus and superinfection of the infected cell population. Once a chronic state is established, the level of linear unintegrated DNA in the surviving cells decreases ~100-fold and is accompanied by the stable integration of several copies of the provirus. It is therefore interesting that similar preliminary observations have been made for LAV/HTLV-III infection of T lymphocytes. Southern blot analysis of DNA from

the chronically infected H9/HTLV-III T-cell line reveals 5–10 copies of integrated provirus, although there is little evidence for the persistence of more than a small amount of linear unintegrated DNA. By contrast, acute infection of human peripheral lymphocyte cultures with LAV leads to the apparent accumulation of >400 copies per cell of linear unintegrated DNA. Although H9/HTLV-III cells are relatively resistant to the cytopathic effects of HTLV-III[10], the rapid disappearance of virus-producing cells in infected primary lymphocyte cultures suggests that LAV/HTLV-III has a severe cytopathic effect upon the natural host target cell ($OKT4^+$)[8,11,14]. The level of linear unintegrated DNA in H9/HTLV-III cells and infected lymphocyte cultures thus seems to correlate with the relative cytopathic effects of the virus on these cells.

The H9/HTLV-III cell line was established by infection with material from several AIDS patients[10] and therefore may contain sequences corresponding to different viral isolates. The distribution of nucleotide differences between the proviral and cDNA sequences described here must be interpreted with caution as we do not know the exact relationship of a particular clone to any of the multiple provirus copies present in the H9/HTLV-III cell line (Fig. 2*A*). Nevertheless, the independent origin of some of the clones is suggested by the large number of nucleotide differences present (Fig. 7*a*). Particularly evident is a high degree of polymorphism in the *env* coding region (24 of 2,568 nucleotides are different) and the intercistronic region (15/602). Eighteen of the changes in *env* lead to amino acid substitutions, mostly nonconservative. Two changes result in new potential sites of asparagine-linked glycosylation while one change removes such a site and therefore may be especially likely to have significant effects on protein antigenicity. Should the proclivity for change with the HTLV-III envelope seen here reflect a high rate of mutation *in vivo*, this may have significant implications for the ability of the immune system to respond effectively to the virus as well as for the design of vaccines.

We thank Dr Jerome Groopman for providing serum samples, Dr Larry Arthur for providing the $H9/HTLV\text{-}III_B$ cell line established by Dr Robert Gallo; Dr John Bell for communicating unpublished data; Parkash Jurhani and Peter Ng for DNA synthesis, Dr Frances Boches for assistance with the immune blotting experiments; Carol Morita for preparation of the figures and Rebecca Cazares for help in preparation of the manuscript. We also thank Drs Jack Obijeski, David Goeddel, Peter Seeburg and Rey Gomez for support and encouragement and Drs Arthur Levinson, Dennis Kleid and David Martin for comments on the manuscript.

Received 27 December 1984; accepted 22 January 1985.

1. Gottlieb, M. *et al. New Engl. J. Med.* **305,** 1425–1431 (1981).
2. Masur, H. *et al. New Engl. J. Med.* **305,** 1431–1438 (1981).
3. Siegal, F. *et al. New Engl. J. Med.* **305,** 1439–1444 (1981).
4. Seligman, M. *et al. New Engl. J. Med.* **311,** 1286–1292 (1984).
5. Lane, H., Masur, H., Edgar, G., Whalen, G. & Fauci, A. *New Engl. J. Med.* **309,** 453–458 (1983).
6. Ziegler, J. *et al. New Engl. J. Med.* **311,** 565–570 (1984).
7. Barre-Sinoussi, F. *et al. Science* **220,** 868 (1983).
8. Montagnier, L. *et al. Human T-Cell Leukemia Viruses,* 363–379 (Cold Spring Harbor Laboratory, New York, 1984).
9. Vilmer, E. *et al. Lancet* **i,** 753–757 (1984).
10. Popovic, M., Sarngadharan, M., Read, E. & Gallo, R. *Science* **224,** 497–500 (1984).
11. Gallo, R. *et al. Science* **224,** 500–503 (1984).
12. Feorino, P. *et al. Science* **225,** 69–72 (1984).
13. Levy, J. *et al. Science* **225,** 840–842 (1984).
14. Klatzmann, D. *et al. Science* **225,** 59–63 (1984).
15. Schupbach, J. *et al. Science* **224,** 503–505 (1984).
16. Sarngadharan, M., Popovic, M., Bruch, L., Schupbach, J. & Gallo, R. *Science* **224,** 505–508 (1984).
17. Safai, B. *et al. Lancet* **i,** 1438–1440 (1984).
18. Brun-Vezinet, F. *et al. Lancet* **i,** 1253–1256 (1984).
19. Brun-Vezinet, F. *et al. Science* **226,** 453–456 (1984).
20. Goedert, J. *et al. Lancet* **ii,** 711–715 (1984).
21. Lawrence, J. *et al. New Engl. J. Med.* **311,** 1269–1273 (1984).
22. Schwartz, D., Zamecnik, P. & Weith, H. *Proc. natn. Acad. Sci. U.S.A.* **74,** 994–998 (1977).
23. Haseltine, W., Maxam, A. & Gilbert, W. *Proc. natn. Acad. Sci. U.S.A.* **74,** 989–993 (1977).
24. Shine, J. *et al. Proc. natn. Acad. Sci. U.S.A.* **74,** 1473–1477 (1977).
25. Varmus, H. *Science* **216,** 812–820 (1982).
26. Peters, G. & Glover, C. *J. Virol.* **35,** 31–40 (1980).
27. Seiki, M., Hattori, S., Hirayama, Y. & Yoshida, M. *Proc. natn. Acad. Sci. U.S.A.* **80,** 3618–3622 (1983).
28. Proudfoot, N. & Brownlee, G. *Nature* **252,** 359–362 (1974).
29. Grosschedl, R. & Birnsteil, M. *Proc. natn. Acad. Sci. U.S.A.* **77,** 1432–1436 (1980).
30. Kitchen, L. *et al. Nature* **312,** 367–369 (1984).
31. Shinnick, T., Lerner, R. & Sutcliffe, J. *Nature* **293,** 543–548 (1981).
32. Schwartz, D., Tizard, R. & Gilbert, W. *Cell* **32,** 853–869 (1983).
33. Bolognesi, D., Luftig, R. & Schafer, J. *Virology* **56,** 549–564 (1973).
34. Smith, B. & Bailey, J. *Nucleic Acids Res.* **7,** 2055–2072 (1979).
35. Arya, S. *et al. Science* **225,** 927–930 (1984).
36. Hahn, B. *et al. Nature* **312,** 166–170 (1984).
37. Majors, J. & Varmus, H. *J. Virol.* **47,** 495–504 (1983).
38. Perlman, D. & Halvorson, H. *J. molec. Biol.* **167,** 394–409 (1983).
39. Montagnier, L. L. *et al. Science* **225,** 63–66 (1984).
40. Haseltine, W. *et al. Science* **225,** 419–421 (1984).
41. Shimotohno, K. *et al. Proc. natn. Acad. Sci. U.S.A.* **81,** 6657–6661 (1984).
42. Weiss, S., Varmus, H. & Bishop, M. *Cell* **12,** 983–992 (1977).
43. Mellon, P. & Duesberg, P. *Nature* **270,** 631–634 (1977).
44. Cordell, B., Weiss, S., Varmus, H. & Bishop, M. *Cell* **15,** 79–91 (1978).
45. Weiss, R., Teich, N., Varmus, H. & Coffin, J. *RNA Tumor Viruses* (Cold Spring Harbor Laboratory, New York, 1982).
46. Popovic, M. *et al. Science* **219,** 856 (1983).
47. Sodrowski, J., Rosen, C. & Haseltine, W. *Science* **225,** 381–385 (1984).
48. Chen, I., McLaughlin, J. & Golde, D. *Nature* **309,** 276–279 (1984).
49. Keshet, E. & Temin, H. *J. Virol.* **31,** 376–388 (1979).
50. Weller, S., Joy, A. & Temin, H. *J. Virol.* **33,** 494–506 (1980).
51. Temin, H., Keshet, E. & Weller, S. *Cold Spring Harbor Symp. quant. Biol.* **44,** 773–777 (1980).
52. Capon, D. *et al. Nature* **304,** 507–513 (1983).
53. Wood, W. *et al. Nature* **312,** 330–337 (1984).
54. Hyunh, T., Young, R. & Davis, R. in *Practical Approaches to Biochemistry* (ed. Groves, D.) (IRL, Oxford, in the press).
55. Benton, W. & Davis, R. *Science* **196,** 180–182 (1977).
56. Foley, G. *et al. Cancer* **18,** 522–529 (1965).
57. Mullins, J., Brody, D., Binari, R. & Cotter, S. *Nature* **308,** 856–858 (1984).
58. Southern, E. *J. molec. Biol.* **98,** 503–517 (1975).
59. Taylor, J., Illmensee, R. & Summer, S. *Biochim. biophys. Acta* **442,** 324–330 (1976).
60. Maniatis, T., Fritsch, E. & Sambrook, J. *Molecular Cloning* (Cold Spring Harbor Laboratory, New York, 1982).
61. Messing, J., Crea, R. & Seeburg, P. *Nucleic Acids Res.* **9,** 309–321 (1981).
62. Sanger, F. *et al. J. molec. Biol.* **143,** 161–178 (1980).
63. Kyte, J. & Doolittle, R. *J. molec. Biol.* **157,** 105–132 (1982).

A *trans*-acting factor is responsible for the simian virus 40 enhancer activity *in vitro*

Paolo Sassone-Corsi, Alan Wildeman & Pierre Chambon

Laboratoire de Génétique Moléculaire des Eucaryotes du CNRS, Unité 184 de Biologie Moléculaire et de Génie Génétique de l'INSERM, Faculté de Médecine, 11 Rue Humann, 67085 Strasbourg Cédex, France

Stimulation of in vitro *transcription by the simian virus 40 enhancer involves a rapid and stable binding of a* trans*-acting factor with both the 5'- and 3'-domains of the enhancer sequence. The enhancer factor, which differs from other types of transcriptional factors, can interact with other enhancer elements.*

REGULATION of gene expression at the level of transcription is probably an important control mechanism during development and in the terminally differentiated cells of eukaryotic organisms. This control may result from the interaction between specific DNA sequences, regulatory proteins and the transcriptional machinery[1]. The promoter DNA sequences involved in the control of transcription initiation in eukaryotic protein-coding genes are composed of several elements: the mRNA start-site, the TATA box sequence and one or several upstream elements, located generally in ~110 base pairs (bp) upstream