JOURNAL OF VIROLOGY, July 1985, p. 101–110
0022-538X/85/070101-10$02.00/0

Vol. 55, No. 1

mRNA Sequence of Three Respiratory Syncytial Virus Genes Encoding Two Nonstructural Proteins and a 22K Structural Protein

NARAYANASAMY ELANGO,† MASANOBU SATAKE,‡ AND SUNDARARAJAN VENKATESAN§*

Laboratory of Infectious Diseases, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland 20205

Received 14 January 1985/Accepted 27 March 1985

An mRNA sequence of two human respiratory syncytial viral nonstructural protein genes and of a gene for a 22,000-molecular-weight (22K) protein was obtained by cDNA cloning and DNA sequencing. Sequences corresponding to the 5′ ends of the respective transcripts were deduced directly by primer extension and dideoxy nucleotide sequencing of the mRNAs. The availability of a bicistronic clone (pRSC$_6$) confirmed the gene order for this portion of the genome. Contrary to other unsegmented negative-stranded RNA viruses, a 19-nucleotide intercistronic sequence was present between the NS$_1$ and NS$_2$ genes. The translation of cloned viral sequences in the bicistronic and monocistronic clones (pRSNS$_1$ and pRSNS$_2$) revealed two moderately hydrophobic proteins of 15,568 and 14,703 daltons. Their similarity in molecular size explained our earlier inability to resolve these proteins. A DNA sequence of an additional recombinant plasmid (pRSA$_2$) revealed a long open reading frame encoding a 22,156-dalton protein containing 194 amino acids. It was relatively basic and moderately hydrophobic. A protein of this size was readily translated in vitro from a viral mRNA hybrid selected by this plasmid and corresponded to an unglycosylated 22K protein seen in purified extracellular virus but not associated with detergent- and salt-resistant cores. A second open reading frame of 90 amino acids partially overlapping with the C terminus of the 22K protein was also present within this sequence. This was reminiscent of the viral matrix protein gene which was previously shown by us to contain two overlapping reading frames. The finding of three additional viral transcripts encoding at least three identifiable proteins in human respiratory syncytial virus was a novel departure from the usual genetic organization of paramyxoviruses. The 5′ ends of all three transcripts had a 5′NGGGCAAAU sequence that is common to all viral transcripts analyzed so far. Although there was no obvious homology immediately upstream of the polyadenylate tail, an AGUUA (AGUAA in the case of NS$_2$) was present between 1 and 4 nucleotides upstream of the polyadenylate end of NS$_1$ and 22K protein mRNAs.

Human respiratory syncytial (RS) virus is a cytoplasmically replicating RNA virus of negative polarity (15). It has been shown to encode at least 10 polyadenylate [poly(A)]-containing mRNAs. Gene coding assignments based on cell-free translation of individual mRNAs have identified protein products for nine of these species (2, 14, 16). Analyses of infected cells and extracellular virus have revealed three major proteins integrated in the viral envelope, namely, an 84,000-molecular-weight (84K) envelope glycoprotein, a 68K fusion glycoprotein, and a 28K nonglycosylated membrane or matrix (M) protein. Detergent-solubilized viral cores are composed of a 46K nucleocapsid (NC) protein, a 27K phosphoprotein (P), and a large protein of ca. 200,000 molecular weight (presumably the viral polymerase) (1, 8, 21, 26, 38). In addition, two or three small proteins (9.5K to 15K) are visualized in infected cell lysates but are not usually present in purified virus and are hence referred to as nonstructural (NS) proteins (2, 16, 38). Unlike the paramyxovirus NS proteins, RS virus apparently encodes the NS proteins within separate genetic units (2, 38). In addition to the above proteins, a 24K envelope-associated protein and its putative mRNA have been recently identified (14, 21). Based on UV kinetics of inactivation of viral transcription (7) and analysis of polycistronic viral transcripts (4), a novel genetic map has been proposed which places two viral NS protein genes rather than NC at the 3′ end of the genome. These data suggest that RS virus has a far more complex genetic organization than do the well-studied paramyxovirus or rhabdovirus prototypes.

Previously we reported the preparation and characterization of cDNA clones representing seven viral mRNAs (38, 39). One of the viral recombinants (pRSC$_6$), which possessed a cDNA insert of ca. 1,050 base pairs (bp), reacted with a viral mRNA of ca. 600 nucleotides, raising the possibility that adventitious viral sequences present in this clone might have resulted from fortuitious cloning of aberrant transcripts. The initial DNA sequence of the viral insert revealed two tandem nonoverlapping open reading frames (ORFs) of ca. 500 bases each, suggesting that a cDNA copy of a bicistronic transcript had been cloned (39). Cell-free translation of a viral mRNA(s) hybrid selected by the recombinant, however, yielded a single polypeptide on polyacrylamide gel electrophoresis under reducing conditions (38). An additional RS virus cDNA clone (pRSA$_2$) possessing a ca. 1,000-bp insert and reacting with a distinct viral RNA species of similar size was also identified (39). Although a translation product for this mRNA was not readily detected, preliminary characterization of detergent-solubilized viral proteins suggested that it might be the candidate mRNA for a 22K unglycosylated protein. To resolve these questions,

* Corresponding author.

† Present address: Laboratory of Viral Diseases, National Institute of Allergy and Infectious Diseases, Bethesda, MD 20205.

‡ Present address: Basic Research Program, Litton Bionetics Inc., Frederick Cancer Research Facility, Frederick, MD 21701.

§ Present address: Office of the Scientific Director, National Institute of Allergy and Infectious Diseases, Frederick Cancer Research Facility, Frederick, MD 21701.

we have undertaken a systematic analysis of these viral genes. Recently, we reported the amino acid sequence of three viral proteins, namely, NC, M, and P, deduced from the DNA sequence of the recombinant plasmids (9, 32, 33). To these we now add the sequence of three more viral genes encoding two NS proteins (NS_1 and NS_2) and a 22K structural protein.

MATERIALS AND METHODS

mRNA isolation. RS virus strain A2 was used to infect HEp-2 cell monolayers. Techniques relating to mRNA isolation and cell-free translation of RNA have been previously described (38). Actinomycin D (1 μg/ml) was routinely used to suppress cellular transcription.

DNA sequencing. Initial characterization of selected cDNA plasmids encoding seven viral genes has been previously documented (38, 39). Nonoverlapping plasmids were segregated on the basis of specific hybridization to distinct viral mRNAs (39). Restriction fragments representing cloned viral sequences of individual recombinants were labeled by nick translation and used to recover additional unique clones from the library by colony hybridization (36). All DNA sequencing was as described by Maxam and Gilbert (24). Computer analysis of DNA sequences was done with the Queen and Korn program (28). Homology comparisons among different proteins were done with the algorithm developed by Wilbur and Lipman (40), and hydropathicity determinations were as described by Kyte and Doolittle (18).

Hybrid selection of viral mRNAs and 5′-end sequencing of individual transcripts. Plasmid DNA (ca. 40 μg) was digested partially with *Hpa*II (0.1 U/μg of DNA, 30 min, 37°C), deproteinized by buffer-saturated phenol-chloroform (1:1) extraction, and denatured by heating to 95°C for 5 min in 0.1 N NaOH. After neutralization, DNA was immobilized to nitrocellulose filters (2.5 cm^2) in 6× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate) and used for hybrid selection experiments. Conditions for hybrid selection of mRNAs and cell-free translation of selected RNAs in a messenger-dependent rabbit reticulocyte translation system were as described previously (33).

For primer extensions, appropriate DNA restriction fragments (1 pmol) labeled at the 5′ end by polynucleotide kinase and [γ-^{32}P]ATP were hybridized to poly(A) RNA from infected cells in buffer containing 80% formamide, 40 mM Na-PIPES [piperazine-*N*,*N*′-bis(2-ethanesulfonic acid] (pH 6.2), 0.4 M NaCl, and 0.2% sodium dodecyl sulfate, for 4 to 16 h at 42°C. RNA-DNA hybrids were recovered, and primer was extended on the RNA template as previously described (33). Input RNA concentrations were varied to obtain almost quantitative conversion of the labeled primer. Dideoxy sequencing of the mRNA 5′ ends has been previously described (33).

RESULTS

Translation of two viral NS proteins. Seventy-five recombinant plasmids were initially selected from an RS cDNA library on the basis of their hybridization to ^{32}P-labeled mRNA from infected cells treated with actinomycin-D, end-labeled viral genomic RNA, and ^{32}P-labeled single-stranded cDNA synthesized in vitro with viral mRNA (38, 39). The plasmids were subsequently segregated into nonoverlapping classes by dot blot cross-hybridization. Plasmids with cDNA inserts possessing isomorphous restriction cleavage patterns were grown together and used to select mRNA from infected cells that was translated in vitro in a rabbit reticulocyte translation system. Serial hybrid selections with individual plasmids allowed us to positively identify plasmids encoding the NC, P, and M proteins (38). Other viral plasmids that were unable to select translatable RNAs were segregated on the basis of their hybridization to distinct viral RNA species resolved by formaldehyde-agarose gel electrophoresis (39). Two viral plasmids ($pRSB_8$ and $pRSC_6$) containing cDNA inserts of 850 and 1,050 bp, respectively, were deemed to encode a viral NS protein gene based on cell-free translation of hybrid-selected viral mRNA (38). $pRSC_6$ possessed a long poly(A) tract and had extensive sequence homology with $pRSB_8$ at one end, corresponding to the putative 5′ end of the viral transcript. Also, $pRSB_8$ lacked the poly(A) tail and several upstream residues found in $pRSC_6$. Therefore, $pRSC_6$ was selected for complete sequence determination. Translation of the cloned DNA sequence revealed two tandem nonoverlapping ORFs potentially encoding polypeptides of 124 and 139 amino acids (Fig. 1). The calculated molecular weight of the encoded proteins was similar to the 15K protein translated from the hybrid-selected mRNA.

To test the possibility that $pRSC_6$ had a cDNA copy of tandemly linked transcript, separate restriction fragments lying within the two reading frames were labeled in vitro by nick translation and used to recover additional recombinants corresponding to each of the two reading frames. Two nonoverlapping clones ($pRSNS_1$ and $pRSNS_2$) containing ca. 500 bp of viral sequence were thus identified. Plasmid DNAs from these two clones as well as $pRSC_6$ were then immobilized to nitrocellulose and used to select mRNA from infected cells for cell-free translation. Two polypeptides with apparent molecular weights of ca. 15,000 and 14,000 (NS_1 and NS_2) were translated from the RNA selected by $pRSC_6$ (Fig. 2, lane B). In contrast, either the 14K (NS_2) or the 15K (NS_1) polypeptide was conspicuously absent among the translation products of RNA selected by $pRSNS_1$ or $pRSNS_2$ (lanes A and C), suggesting the presence of two distinct viral mRNAs.

Evidence for two transcriptional units. Poly(A)-containing RNAs from infected or uninfected cells were resolved by

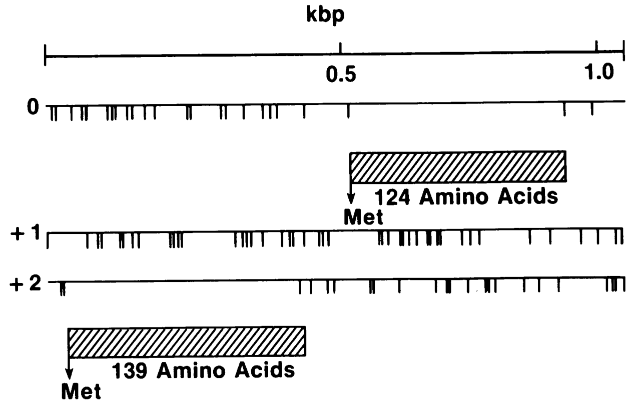

FIG. 1. Translation of the messenger strand of the cloned viral insert in $pRSC_6$ in three reading frames. The length in kilobase pairs (kbp) of the cloned sequence is indicated at the top. The two long ORFs are emphasized by the shaded rectangles, with the number of amino acids within them shown underneath. The small vertical lines denote the stop codons.

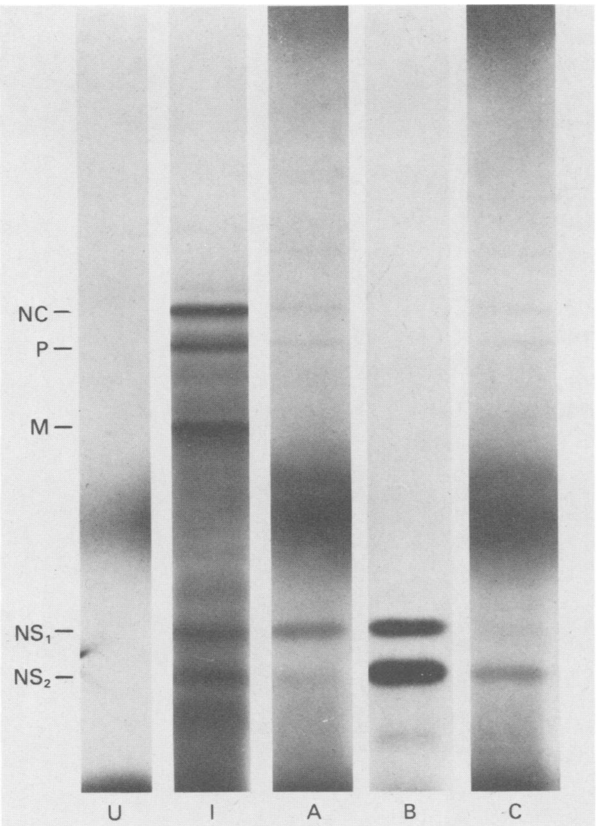

FIG. 2. Cell-free translation of hybrid-selected RS viral mRNAs. Poly(A)-containing RNAs from uninfected (lane U) or infected (lane I) cells treated with actinomycin D were translated in a messenger-dependent rabbit reticulocyte system, and the translation products were resolved by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (15% acrylamide; 150:1 acrylamide-bisacrylamide concentration) under reducing conditions. Lanes A to C, Results of translation of mRNA hybrid selected with recombinant plasmids pRSNS$_1$, pRSC$_6$, and pRSNS$_2$ respectively. In each case, 40 μg of plasmid DNA was partially digested with *Hpa*II (0.1 U/μg of DNA, 37°C, 30 min) and denatured by boiling briefly in 0.1 N NaOH after deproteinization by buffer-saturated phenol-chloroform (1:1) extraction. DNA solution was then neutralized and immobilized to nitrocellulose filters in 6× SSC. The filters were used to hybrid select specific viral RNA as previously described (33). The optimal amount of RNA used for hybridization [ca. 10 μg of poly(A) RNA per 50 μg of DNA] was determined in preliminary experiments designed to obtain DNA saturation kinetics (see legend to Fig. 4). The viral proteins translated from total viral mRNA are identified on the left.

formaldehyde agarose gel electrophoresis and transferred to nitrocellulose filters (37). The RNA filter blots were hybridized separately to ^{32}P-labeled viral inserts of pRSNS$_1$, pRSNS$_2$ and pRSC$_6$. A prominant viral RNA species of ca. 550 bases hybridized to all three probes. Additional minor species of higher-molecular-weight RNA also reacted with the three DNAs to a variable extent. The pattern of the polycistronic transcripts hybridizing to pRSNS$_1$ was different from that reacting with pRSNS$_2$. A similar differential hybridization pattern was also observed by Collins and Wertz, with cloned DNA probes representing two adjacent transcriptional units (4).

To confirm the existence of two discrete mRNAs and precisely locate the transcriptional start site(s), restriction fragments lying downstream of the N termini of the two reading frames in pRSC$_6$ were isolated, labeled at the 5′ end of the antimessenger strands, and annealed to poly(A)-containing RNA from infected cells (Fig. 3). When either a 32-bp *Rsa*I-*Fnu*4HI fragment labeled at the *Fnu*4HI site or a 44-bp *Nco*I-*Sau*3AI fragment labeled at the *Sau*3AI site was used during a reverse-transcriptase-catalyzed primer extension reaction, prominent cDNA products of 60 or 75 nucleotides were synthesized (data not shown). No extension occurred if the 5′ labels were on the opposite strand, thus confirming the transcriptional polarity within cloned DNA. The *Nco*I-*Sau*3AI fragment was not extended significantly beyond 30 nucleotides, thus eliminating the predominant presence of a long contiguous mRNA containing both NS$_1$ and NS$_2$ coding sequences. Partial reverse-transcriptase-catalyzed reactions were then done with the same primers and dideoxynucleoside 5′-triphosphate inhibitors to obtain sequences complementary to 5′ ends of the respective mRNAs. Figure 3 shows the 5′-end sequences of the transcripts deduced in this manner. Starting at the penultimate nucleotide from the 5′ end, there was a CCCGUUUAU sequence (antimessenger sense) common to both transcripts, beyond which there was considerable sequence divergence. The first seven nucleotides corresponding to the 5′ end of the proximal transcript (NS$_1$) were absent in pRSC$_6$ which, however, contained the transcriptional sequences for both NS$_1$ and NS$_2$ genes and a 23-bp intergenic sequence (see below).

mRNA sequence of two adjacent viral genes. cDNA inserts of pRSNS$_1$ and pRSNS$_2$ were also sequenced. Each had a poly(A) tract at one end and a CAAAU sequence after a chain of G residues at the 5′ end of the transcripts. The cluster of G residues either included those of full-length cDNA or was derived during oligodeoxycytidylate addition of the first strand during cDNA construction. The DNA sequence of the pRSC$_6$ insert is presented in the messenger sense in Fig. 4. The 3′ end of the proximal gene (NS$_1$) (indicated by an arrow) represents the sequence immediately upstream of the poly(A) tail in pRSNS$_1$. The four A residues after this sequence might be the counterparts of four U residues in the genomic RNA. By analogy with other unsegmented negative-stranded RNA viruses, these might be reiteratively copied to generate the poly(A) tail of the mRNA (12). After the four A residues there was a 19-nucleotide sequence before the initiation site for the second (NS$_2$) transcript. This span of 19 nucleotides not included in the mRNA was interpreted to be the intercistronic region on the genome. The first nine nucleotides which constitute the 5′ end of the NS$_2$ transcript are conserved in all RS viral transcripts analyzed by primer extension reactions (3; S. Venkatesan, N. Elango, and M. Satake, unpublished data). Of these, the identity of the first nucleotide as G is not absolute since during primer extension, reverse transcriptase has been known to copy the guanylate residue of the cap structure at the 5′ end of mRNAs (12).

The calculated molecular sizes of 15,568 and 14,703 daltons for the two encoded proteins explain our earlier inability to resolve these two proteins (38, 39). These values were somewhat at variance with the 14K and 11K translation products of two adjacent transcripts reported by Collins et al. (3) and our previous estimates. The deduced amino acid sequences of both of these proteins were relatively unremarkable but for the fact that they were moderately hydrophobic (33.6% for NS$_1$ and 32.9% for NS$_2$), and the NS$_1$ was relatively more acidic. There was no obvious clustering of hydrophobic residues in either protein.

Cell-free translation of a 22K viral protein. During the

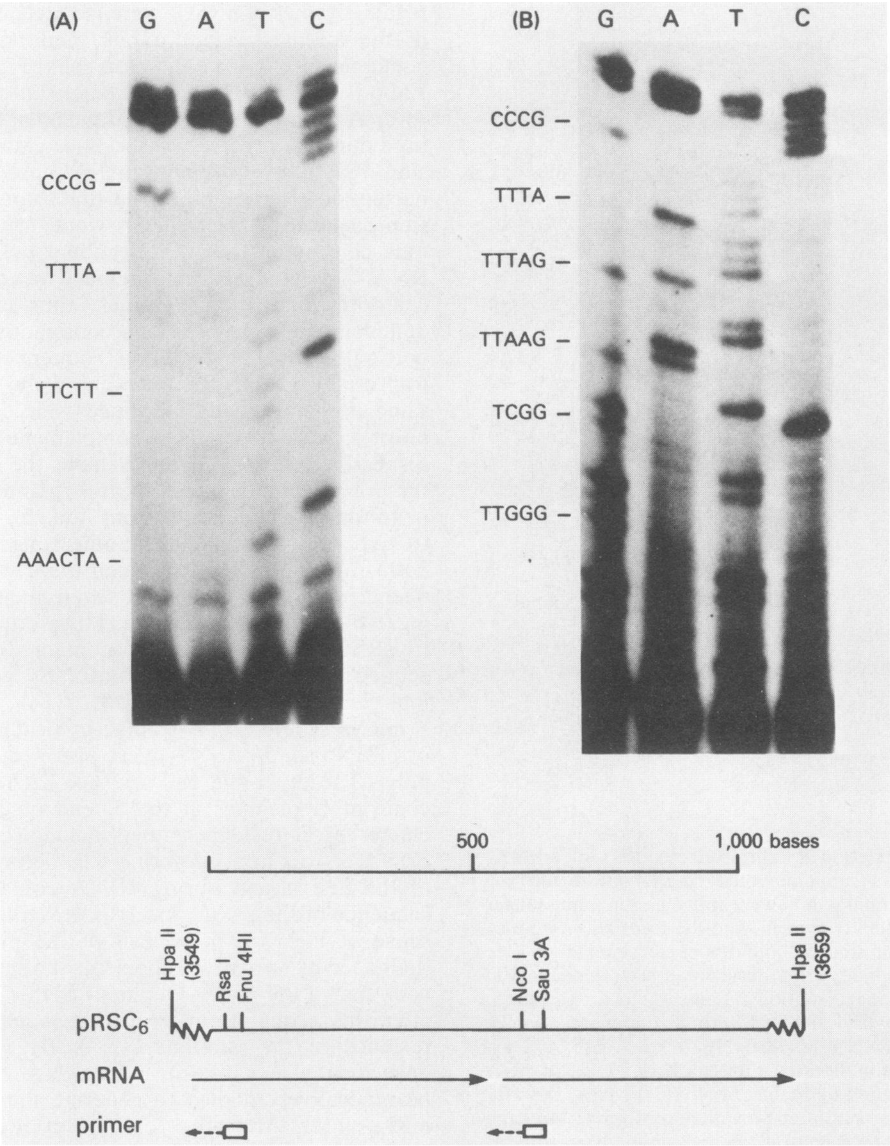

FIG. 3. Sequence of 5′ ends of two viral mRNAs determined by primer extension and dideoxy sequencing. The approximate map coordinates of the viral mRNAs and the strategy used for mapping the 5′ ends are schematically illustrated. The solid line denotes the RS viral sequence within pRSC$_6$, with the flanking pBR322 sequences indicated by wavy extensions. The two *Hpa*II sites occur in pBR322 at the map positions indicated. One picomole each of the 32-bp *Rsa*I-*Fnu*4HI or the 44-bp *Nco*I-*Sau*3AI fragments 5′ end labeled at the *Fnu*4HI and *Sau*3AI site was hybridized to varying concentrations of poly(A) RNA from infected cells. DNA-RNA hybrids were recovered, and the primer was extended in a complete reaction (33) to determine the concentration of RNA required to achieve greater then 50% conversion of primer(s) to the extended products. Partial DNA biosynthetic reactions were carried out by using dideoxynucleoside triphosphate inhibitors as previously described (33). The reaction products were resolved on thin 8% acrylamide urea (6 M) gels. A and B, Sequencing profiles obtained with the *Rsa*I-*Fnu*4HI and *Nco*I-*Sau*3AI primers respectively. Lanes G, A, T, and C are the results when ddGTP, ddATP, ddTTP, and ddCTP, respectively, are present in the reaction. The sequence in each case in negative polarity is on the left.

initial characterization of polypeptides associated with extracellular virus, an unglycosylated 22K protein was also observed (14, 26). This particular protein was readily solubilized from the viral core under conditions that extract the envelope glycoproteins and the M protein (14; D. Prevar, and S. Venkatesan, unpublished data). Earlier, we reported that the viral M protein was frequently present as two forms of 28,000 and 26,000 molecular weight which were related to each other by tryptic fingerprinting (33, 38). This behavior was particularly accentuated when the M protein was translated in vitro (3'2). Initially, we thought the 22K protein in the detergent-solubilized viral extract to be similarly related to the M protein generated by proteolytic artifacts during viral extraction. Subsequent biochemical analysis of an RS cDNA library revealed the presence of a distinct nonoverlapping cDNA recombinant (pRSA$_2$) hybridizing solely to a viral mRNA of ca. 1,050 nucleotides (39). An RNA of similar size designated 3b was shown by Collins et al. to be translated in vitro, yielding a 24K polypeptide (2). pRSA$_2$ was used to select mRNA from infected cells which was then translated in a messenger-dependent rabbit reticulocyte system. A single polypeptide of 22,000 molecular

```
                                                                              5
ATAAGAA   TTTGATAAGT   ACCACTTAAA   TTTAACTCCC   TTGGTTAGAG  ATG GGC AGC AAT TCA
                                                            MET GLY SER ASN SER
                                     15                                      25
TTG AGT ATG ATA AAA GTT AGA TTA CAA AAT TTG TTT GAC AAT GAT GAA GTA GCA TTG TTA
LEU SER MET ILE LYS VAL ARG LEU GLN ASN LEU PHE ASP ASN ASP GLU VAL ALA LEU LEU
                                     35                                      45
AAA ATA ACA TGC TAT ACT GAT AAA TTA ATA CAT TTA ACT AAT GCT TTG GCT AAG GCA GTG
LYS ILE THR CYS TYR THR ASP LYS LEU ILE HIS LEU THR ASN ALA LEU ALA LYS ALA VAL
                                     55                                      65
ATA CAT ACA ATC AAA TTG AAT GGC ATT GTG TTT GTG CAT GTT ATT ACA AGT AGT GAT ATT
ILE HIS THR ILE LYS LEU ASN GLY ILE VAL PHE VAL HIS VAL ILE THR SER SER ASP ILE
                                     75                                      85
TGC CCT AAT AAT AAT ATT GTA GTA AAA TCC AAT TTC ACA ACA ATG CCA GTA CTA CAA AAT
CYS PRO ASN ASN ASN ILE VAL VAL LYS SER ASN PHE THR THR MET PRO VAL LEU GLN ASN
                                     95                                     105
GGA GGT TAT ATA TGG GAA ATG ATG GAA TTA ACA CAT TGC TCT CAA CCT AAT GGT CTA CTA
GLY GLY TYR ILE TRP GLU MET MET GLU LEU THR HIS CYS SER GLN PRO ASN GLY LEU LEU
                                    115                                     125
GAT GAC AAT TGT GAA ATT AAA TTC TCC AAA AAA CTA AGT GAT TCA ACA ATG ACC AAT TAT
ASP ASP ASN CYS GLU ILE LYS PHE SER LYS LYS LEU SER ASP SER THR MET THR ASN TYR
                                    135
ATG AAT CAA TTA TCT GAA TTA CTT GGA TTT GAT CTT AAT CCA TAA     ATTATAATTA
MET ASN GLN LEU SER GLU LEU LEU GLY PHE ASP LEU ASN PRO END
                                                       ↓
ATATCAACTA   GCAAATCAAT   GTCACTAACA   CCATTAGTTA   ATATAAAACT   TAACAGAAGA
                                                                              8
CAAAAATGGG   GCAAATAAAT   CAATTCAGCC  AACCCAACC ATG GAC ACA ACC CAC AAT GAT AAT
                                                MET ASP THR THR HIS ASN ASP ASN
                                     18                                      28
ACA CCA CAA AGA CTG ATG ATC ACA GAC ATG AGA CCG TTG TCA CTT GAG ACC ATA ATA ACA
THR PRO GLN ARG LEU MET ILE THR ASP MET ARG PRO LEU SER LEU GLU THR ILE ILE THR
                                     38                                      48
TCA CTA ACC AGA GAC ATC ATA ACA CAC AAA TTT ATA TAC TTG ATA AAT CAT GAA TGC ATA
SER LEU THR ARG ASP ILE ILE THR HIS LYS PHE ILE TYR LEU ILE ASN HIS GLU CYS ILE
                                     58                                      68
GTG AGA AAA CTT GAT GAA AGA CAG GCC ACA TTT ACA TTC CTA GTC AAC TAT GAA ATG AAA
VAL ARG LYS LEU ASP GLU ARG GLN ALA THR PHE THR PHE LEU VAL ASN TYR GLU MET LYS
                                     78                                      88
CTA TTA CAC AAA GTA GGA AGC ACT AAA TAT AAA AAA TAT ACT GAA TAC AAC ACA AAA TAT
LEU LEU HIS LYS VAL GLY SER THR LYS TYR LYS LYS TYR THR GLU TYR ASN THR LYS TYR
                                     98                                     108
GGC ACT TTC CCT ATG CCA ATA TTC ATC AAT CAT GAT GGG TTC TTA GAA TGC ATT GGC ATT
GLY THR PHE PRO MET PRO ILE PHE ILE ASN HIS ASP GLY PHE LEU GLU CYS ILE GLY ILE
                                    118
AAG CCT ACA AAG CAT ACT CCC ATA ATA TAC AAG TAT GAT CTC AAT CCA TAA ATTTCAACAC
LYS PRO THR LYS HIS THR PRO ILE ILE TYR LYS TYR ASP LEU ASN PRO END

AATATTCACA CAATCTAAAA CAACAACTCT ATGCATAACT ATACTCCATA GTCCAGATGG AGCCTGAAAA

TTATAGTAAT TTAAAAAAAA AAAAA
```

MOLECULAR WEIGHT OF NS1 = 15568

MOLECULAR WEIGHT OF NS2 = 14703

FIG. 4. DNA sequence of two tandem RS viral genes and the deduced amino acid sequence of the encoded proteins. The DNA sequence presented in the messenger sense represents the entire viral sequence cloned in pRSC$_6$. The sequence in the coding regions for the two proteins was independently confirmed by DNA sequencing of inserts within pRSNS$_1$ and pRSNS$_2$ plasmids containing the proximal and distal genes. The arrow denotes the end of the viral sequence in pRSNS$_1$ not including the poly(A) tail. The boxed-in nonanucleotide GGGGCAAAT sequence was identical to the 5′ end of NS$_2$ transcript deduced by primer extension and dideoxy sequencing (Fig. 3), barring the first G residue. The four A residues after the arrow probably represent the messenger equivalent of the polyadenylation signal. The 19-nucleotide gap between the four A residues and the 5′ end of the NS$_2$ transcript were unique and thought to represent the intercistronic region.

weight was translated only when RNA from infected cells was used for hybrid selection (Fig. 5, lane B).

Sequence of mRNA for the 22K proteins. The recombinant pRSA$_2$ possessed 14 A residues at one end which corresponded to the oligodeoxythymidylate-primed first cDNA strand synthesis. The cDNA insert was completely sequenced by the chemical method of Maxam and Gilbert (24) and used to obtain a complete restriction map. A 78-bp *Hin*fI-*Hae*III fragment lying within the presumptive 5′ region of the mRNA was 5′ end labeled at the *Hae*III site (Fig.

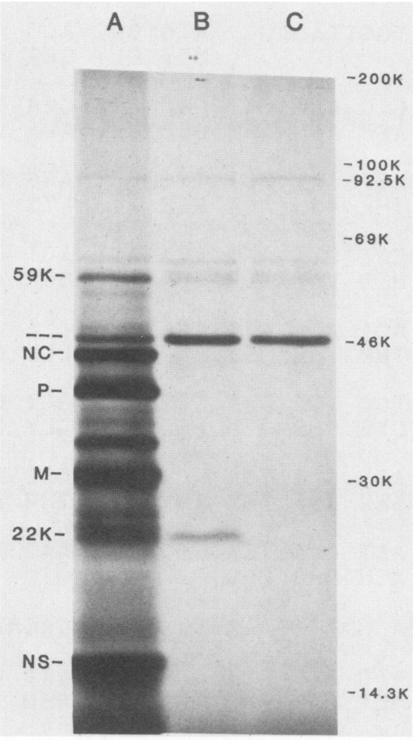

FIG. 5. Cell-free translation of RS viral mRNA hybrid selected by recombinant pRSA$_2$. Conditions for hybridization and cell-free translation have been described in the text. Poly(A)-containing RNA from infected (B) or uninfected (C) cells was selected with pRSA$_2$ plasmid DNA. The translation profile of unselected RNA is shown (A), with the viral proteins identified on the left. The interrupted line denotes the endogenous 46K band seen in reticulocyte translation systems.

6) and annealed to viral mRNA, and the primer was extended with reverse transcriptase. Primer extension occurred only when the label was present at the *Hae*III site, thus confirming the polarity of the message within the cloned insert. Partial DNA biosynthetic reactions with dideoxynucleoside 5′-triphosphate inhibitors were then used to obtain the 5′-end sequence of the message. As shown in Fig. 6, the viral insert was found to lack the NGGG sequence constituting the 5′-proximal portion of the decanucleotide NGGGCAAAU(A) sequence conserved at the 5′ end of all viral transcripts (3; Venkatesan et al., unpublished data).

The DNA sequence in the messenger orientation was translated to yield a 22,156-dalton protein containing 194 amino acids (Fig. 7). The deduced protein was slightly basic (14.8% basic versus 11.8% acidic residues) and moderately hydrophobic (27.1%). It had no homology with any other viral proteins or known eucaryotic cell proteins. There was no obvious clustering of hydrophobic residues, unlike the RS virus M protein (32). The codon usage for this protein as well as the two NS proteins showed an inherent bias against CG dinucleotide similar to other animal viruses and eucaryotic genomes (27, 31, 35). A second ORF potentially encoding a protein of 10,675 daltons and containing 90 amino acids was also present (Fig. 8). This overlapped the C terminus of the longer ORF by 31 amino acids. This genetic organization was analogous to the viral M gene which was shown to potentially encode a polypeptide of 75 amino acids in an alternate reading frame partially overlapping the C terminus of the M protein (32). Viral proteins corresponding to either of these two ORFs were not observed, and it is unlikely that the 9.5K protein described by Collins et al. (2) is related to either of them since the latter was shown to be derived from a discrete mRNA of 0.2 × 10^6 daltons.

DISCUSSION

Molecular studies relating to the genetic organization of RS virus initiated in our laboratory as well as those by Wertz and colleagues have begun to unravel certain novel features peculiar to this virus. Although originally classified as a paramyxovirus, RS virus has been placed in a separate subgroup *Pneumovirus* on the basis of distinct helical morphology of the viral NC and the lack of hemagglutinin and neuraminidase activities (17). The genetic complexity and the organization of this virus are unusual in several respects. (i) Unlike the usual 6 transcripts of paramyxoviruses, the negative-stranded genome of this virus encodes 10 transcripts, with polypeptide products for 9 of them having been previously identified (2, 39). (ii) Two viral NS protein genes rather than NC are 3′ proximal on the genome (2, 4, 7). (iii) Unlike the paramyxoviruses, NS proteins are encoded separate genetic units (2, 39), and there is no equivalent of Sendai virus C protein encoded by a second overlapping reading frame within the P gene (11, 34).

Both of the NS transcripts are initiated with a 5′-NGGGCAAAU sequence that is common to eight viral transcripts examined so far. At the 3′ end of the transcripts, there is either an AGUUA or AGUAA sequence 4 or 3 nucleotides upstream of the poly(A) tail, respectively. The AGUUA sequence is also observed 1 to 4 nucleotides upstream of the poly(A) tract of five other viral transcripts (3; Venkatesan et al., unpublished data). This is somewhat reminiscent of the sequence arrangement at the ends of vesicular stomatitis virus and Sendai virus transcripts (12). The sequence of the bicistronic clone was consistent with the genetic map proposed by Collins et al. which placed the NS$_1$ gene proximal to the NS$_2$ genome. Presently there is no evidence for the existence of an untranslated leader RNA at the 3′ end of the genome of this virus similar to what has been noted for other unsegmented negative-stranded RNA viruses (17). Primer extension experiments designed to determine the 5′-end sequence of NS$_1$ mRNA occasionally revealed minor extensions of ca. 50 nucleotides upstream of the true 5′ end of this transcript (preliminary observations). Whether this represents a transcript linking the leader RNA to NS$_1$ transcript remains to be proven by direct 3′-end sequencing of the genomic RNA. There were 23 nucleotides separating mRNA sequences of the NS$_1$ and NS$_2$ genes. This was unexpected since both vesicular stomatitis virus and Sendai virus have a short (di- or trinucleotide) spacer sequence separating adjacent transcripts. Of the 23 nucleotides, the proximal four U residues (genomic sense) might be the equivalent of the seven or five U residues present in vesicular stomatitis virus or Sendai virus and, presumably, reiteratively copied during transcription to generate the poly(A) tail (12, 30). Interestingly, Paterson et al. have also noted a 22-nucleotide spacer separating M and Fo mRNA sequences in a bicistronic simian virus 5 virus cDNA plasmid (25). Although artifacts resulting from template switching errors during the synthesis of cDNA could not be excluded, if this unusual pattern of long intergenic boundaries prevails with other genes, then it is of some evolutionary interest. The genome of influenza virus has conserved sequences at the 3′ and 5′ ends of all of its RNA segments.

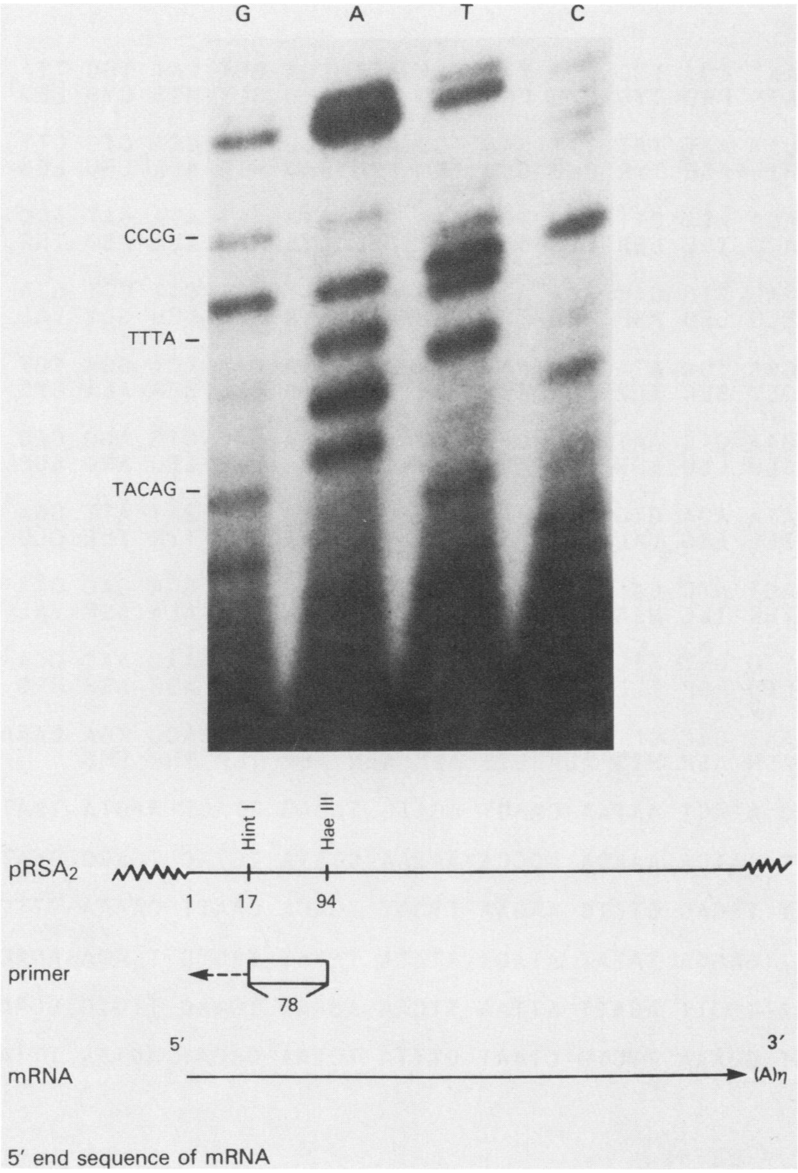

FIG. 6. Primer extension and dideoxy sequencing of mRNA for the 5′ end of the 22K protein. The experimental strategy and the transcriptional coordinates are schematically shown under the sequencing profile. The cloned viral insert is denoted by the straight line, and the wavy lines indicate the pBR322 sequence up to the *Hpa*II sites on either side of the *Pst*I cloning site. Initial DNA sequencing established the messenger polarity within the cloned insert, and the 78-bp *Hin*fI-*Hae*III fragment (5′ end labeled on the negative strand at the *Hae*III site) was used as the primer. The base-specific dideoxy sequencing reactions are shown in lanes G, A, T, and C, with the sequence in the antimessenger sense shown on the left.

Primer-dependent transcription initiates at the penultimate nucleotide of each segment, and the transcripts terminate at a stretch of U residues ca. 17 nucleotides upstream of the 5′ end of the genomes that are not copied into mRNA (19, 29, 30). It can then be speculated that the long intercistronic spacer(s) in RS virus are evolutionary remnants of the nontranscribed 5′ sequences of segmented influenza virus-like ancestor that was tandemly ligated to generate the linear genome.

Several paramyxoviruses have been shown to encode one or two small NS proteins either from a second overlapping ORF within the P gene (Sendai and measles) (11, 34; B. Bellini, personal communication) or from the N-terminal portion of the P protein itself as in Newcastle disease virus (5). Interestingly, the mRNA sequence of RS virus P gene had a single ORF, and distinct transcripts for three viral NS proteins have been identified (2, 16, 32). Although no functional roles have been ascribed to RS viral NS proteins, transcripts for two of these are the earliest to be synthesized, and it might be argued that the protein products of at least two of these genes may be involved in transcriptional modulation. Sendai virus C protein synthesized from the mRNA for P protein apparently is temporally regulated in infected cells at the translational level and is presumed to play a role in viral replication (6, 20).

Although a certain amount of genetic economy has been sacrificed by the incorporation of discrete transcriptional units for specific polypeptides, such an organization in RS

```
CAAAT
                                     10                                      20
ATG TCA CGA AGG AAT CCT TGC AAA TTT GAA ATT CGA GGT CAT TGC TTA AAT GGT AAG AGG
MET SER ARG ARG ASN PRO CYS LYS PHE GLU ILE ARG GLY HIS CYS LEU ASN GLY LYS ARG
                                     30                                      40
TGT CAT TTT AGT CAT AAT TAT TTT GAA TGG CCA CCC CAT GCA CTG CTT GTA AGA CAA AAC
CYS HIS PHE SER HIS ASN TYR PHE GLU TRP PRO PRO HIS ALA LEU LEU VAL ARG GLN ASN
                                     50                                      60
TTT ATG TTA AAC AGA ATA CTT AAG TCT ATG GAT AAA AGT ATA GAT ACC TTA TCA GAA ATA
PHE MET LEU ASN ARG ILE LEU LYS SER MET ASP LYS SER ILE ASP THR LEU SER GLU ILE
                                     70                                      80
AGT GGA GCT GCA GAG TTG GAC AGA ACA GAA GAG TAT GCT CTT GGT GTA GTT GGA GTG CTA
SER GLY ALA ALA GLU LEU ASP ARG THR GLU GLU TYR ALA LEU GLY VAL VAL GLY VAL LEU
                                     90                                     100
GAG AGT TAT ATA GGA TCA ATA AAC AAT ATA ACT AAA CAA TCA GCA TGT GTT GCC ATG AGC
GLU SER TYR ILE GLY SER ILE ASN ASN ILE THR LYS GLN SER ALA CYS VAL ALA MET SER
                                    110                                     120
AAA CTC CTC ACT GAA CTC AAT AGT GAT GAT ATC AAA AAG CTG AGG GAC AAT GAA GAG CTA
LYS LEU LEU THR GLU LEU ASN SER ASP ASP ILE LYS LYS LEU ARG ASP ASN GLU GLU LEU
                                    130                                     140
AAT TCA CCC AAG ATA AGA GTG TAC AAT ACT GTC ATA TCA TAT ATT GAA AGC AAC AGG AAA
ASN SER PRO LYS ILE ARG VAL TYR ASN THR VAL ILE SER TYR ILE GLU SER ASN ARG LYS
                                    150                                     160
AAC AAT AAA CAA ACT ATC CAT CTG TTA AAA AGA TTG CCA GCA GAC GTA TTG AAG AAA ACC
ASN ASN LYS GLN THR ILE HIS LEU LEU LYS ARG LEU PRO ALA ASP VAL LEU LYS LYS THR
                                    170                                     180
ATC AAA AAC ACA TTG GAT ATC CAT AAG AGC ATA ACC ATC AAC AAC CCA AAA GAA TCA ACT
ILE LYS ASN THR LEU ASP ILE HIS LYS SER ILE THR ILE ASN ASN PRO LYS GLU SER THR
                 ** *               190
GTT AGT GAT ACA AAT GAC CAT GCC AAA AAT AAT GAT ACT ACC TGA CAAAT ATCCT
VAL SER ASP THR ASN ASP HIS ALA LYS ASN ASN ASP THR THR END

TGTAG TATAA CTTCC ATACT AATAA CAAGT AGATG TAGAG TTACT ATGTA TAATC AAAAG

AACAC ACTAT ATTTC AATCA AAACA ACCCA AATAA CCATA TGTAC TCACC GAATC AAACA

TTCAA TGAAA TCCAT TGGAC CTCTC AAGAA TTGAT TGACA CAATT CAAAA TTTTC TACAA
                                                         ** *
CATCT AGGTA TTATT GAGGA TATAT ATACA ATATA TATAT TAGTG TCATA ACACT CAATT

CTAAC ACTCA CCACA TCGTT ACATT ATTAA TTCAA ACAAT TCAAG TTGTG GGACA AAATG

GATCC CATTA TTAAT GGAAA TTCTG CTAAT GTTTA TCTAA CCGAT AGTTA TTTAA AAAAA

AAAAA AA

MOLECULAR WEIGHT = 22156
```

FIG. 7. DNA sequence of the cloned viral insert of $pRSA_2$. The nucleotide sequence is presented in the messenger sense, with the amino acid sequence of the long ORF encoding the 22K protein. The asterisks delineate the second overlapping ORF.

virus obviates the need for complex fine tuning of the expression of overlapping genes. If it is presumed that the functions specified by the NS proteins originally existed as separate genetic segments in an influenza virus-like ancestor, then the genetic organization of the present-day RS virus could have resulted from simple end-to-end ligation of the primitive segmented genomes. If, on the contrary, analogous viral functions were encoded within alternate ORFs of preexisting segmented ancestral genes or substituted for by the evolution of new ORFs within preexisting genetic units of unsegmented virus, then a condensed genome such as Sendai virus (11) might emerge. By following this line of reasoning, it is tempting to speculate that RS-like virus represents the earliest step in the evolution of negative-stranded genomes since it has both discrete transcriptional units for all its identifiable gene products and additional overlapping reading frames within two viral genes. This structural feature and the possible occurrence of long intergenic regions might be the results of a primitive end-to-end ligation of segmented genomes. Sequence comparison of the intergenic regions in the genomic RNA among various paramyxoviruses and with those of influenza viruses might provide some clues to understanding this mechanism.

A recombinant cDNA plasmid of an additional viral transcript encoding a 22K protein was identified. A protein of this size was consistently observed in the purified virus; although its exact location in the virus was not known precisely, it was not associated with detergent- and salt-resistant viral cores, nor was it glycosylated. It was moderately hydrophobic, relatively basic, and probably has an architectural function analogous to the viral M protein. Interestingly, a second ORF encoding 90 amino acids and partially overlapping with the C terminus of the 22K proteins was present in this transcript in a manner similar to what was earlier reported for the RS virus M gene (33). The significance of these overlapping reading frames is not clear, since

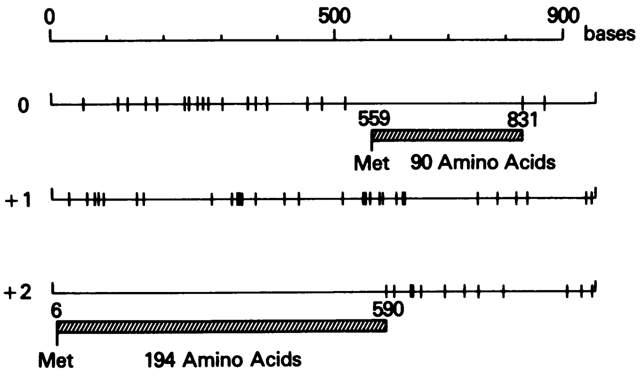

FIG. 8. Schematic translation of RS viral sequence in $pRSA_2$. The length of the cloned insert in bases is shown at the top. The two ORFs are denoted by hatched rectangles with the respective amino acid contents shown underneath. The vertical arrows represent the translation stop codons.

no corresponding viral polypeptides have yet been identified.

ACKNOWLEDGMENTS

We thank Robert M. Chanock for his support and enthusiasm for molecular studies with this virus, Ena Camargo for assistance in virus infection and RNA isolation, and T. Cecil and G. Harbaugh for editorial assistance.

LITERATURE CITED

1. **Bernstein, J. M., and J. F. Hruska.** 1981. Respiratory syncytial virus proteins: identification by immunoprecipitation. J. Virol. **38:**278–285.
2. **Collins, P. L., Y. T. Huang, and G. W. Wertz.** 1984. Identification of a tenth mRNA of respiratory syncytial virus and assignment of polypeptides to the 10 viral genes. J. Virol. **49:**572–578.
3. **Collins, P. L., Y. T. Huang, and G. W. Wertz.** 1984. Nucleotide sequence of the gene encoding the fusion (F) glycoprotein of human respiratory syncytial virus. Proc. Natl. Acad. Sci. U.S.A. **81:**7683–7687.
4. **Collins, P. L., and G. W. Wertz.** 1983. cDNA cloning and transcriptional mapping of nine polyadenylated RNAs encoded by the genome of human respiratory syncytial virus. Proc. Natl. Acad. Sci. U.S.A. **80:**3208–3212.
5. **Collins, P. L., G. W. Wertz, L. A. Ball, and L. E. Hightower.** 1982. Coding assignments of the five smaller mRNAs of Newcastle disease virus. J. Virol. **43:**1024–1031.
6. **Dethlefsen, L., and D. Kolakofsky.** 1983. In vitro synthesis of the nonstructural C protein of Sendai virus. J. Virol. **46:**321–324.
7. **Dickens, L. E., P. L. Collins, and G. W. Wertz.** 1984. Transcriptional mapping of human respiratory syncytial virus. J. Virol. **52:**364–369.
8. **Dubovi, E. J.** 1982. Analysis of proteins synthesized in respiratory syncytial virus-infected cells. J. Virol. **42:**372–378.
9. **Elango, N., and S. Venkatesan.** 1983. Amino acid sequence of respiratory syncytial virus capsid protein. Nucleic Acids Res. **11:**5941–5951.
10. **Fernie, B. F., and J. L. Gerin.** 1982. Immunochemical identification of viral and nonviral proteins of the respiratory syncytial virus virion. Infect. Immun. **37:**243–249.
11. **Giorgi, C., B. M. Blumberg, and D. Kolakofsky.** 1983. Sendai virus contains overlapping genes expressed from a single mRNA. Cell **35:**829–836.
12. **Gupta, K. C., and D. W. Kingsbury.** 1984. Complete sequences of the intergenic and mRNA start signals in the Sendai virus genome: homologies with the genome of vesicular stomatitis virus. Nucleic Acids Res. **12:**3829–3840.
13. **Herman, R. C., M. Shubert, J. D. Keene, and R. A. Lazzarini.** 1980. Polycistronic vesicular stomatitis virus RNA transcripts. Proc. Natl. Acad. Sci. U.S.A. **77:**4662–4665.
14. **Huang, Y. T., Collins, P. L., and G. W. Wertz.** 1984. Identification of a new envelope-associated protein of human respiratory syncytial virus, p. 365–368. *In* D. H. L. Bishop and R. W. Compans (ed.), Nonsegmented negative strand viruses: paramyxoviruses and rhabdoviruses. Academic Press, Orlando, Fla.
15. **Huang, Y. T., and G. W. Wertz.** 1982. The genome of respiratory syncytial virus is a negative-stranded RNA that encodes at least seven mRNA species. J. Virol. **43:**150–157.
16. **Huang, Y. T., and G. W. Wertz.** 1983. Respiratory syncytial virus mRNA coding assignments. J. Virol. **46:**667–672.
17. **Kingsbury, D. W.** 1979. Paramyxoviruses, p. 367–398. *In* D. B. Nayak (ed.), Molecular biology of animal viruses. Marcel Dekker, Inc., New York.
18. **Kyte, J., and R. F. Doolittle.** 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157:**105–132.
19. **Lamb, R. A., and P. W. Choppin.** 1983. The gene structure and replication of influenza virus. Annu. Rev. Biochem. **52:**467–506.
20. **Lamb, R. A., B. W. J. Mahy, and P. W. Choppin.** 1976. The synthesis of Sendai virus polypeptides in infected cells. Virology **69:**116–131.
21. **Lambert, D. M., and M. W. Pons.** 1983. Respiratory syncytial virus glycoproteins. Virology **130:**204–214.
22. **Lehrach, H., D. Diamond, J. Wozney, and H. Boedtker.** 1977. RNA molecular weight determination by gel electrophoresis under denaturing conditions, a critical reexamination. Biochemistry **16:**4743–4751.
23. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
24. **Maxam, A. M., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. **65:**499–560.
25. **Paterson, R. G., T. J. R. Harris, and R. A. Lamb.** 1984. Fusion protein of the paramyxovirus SV5: nucleotide sequence of mRNA predicts a highly hydrophobic glycoprotein. Proc. Natl. Acad. Sci. U.S.A. **81:**6706–6710.
26. **Peeples, M., and S. Levine.** 1979. Respiratory syncytial virus: their location in the virion. Virology **95:**137–145.
27. **Porter, A. G., C. Barber, N. H. Carey, R. A. Hallewall, G. Threlfall, and J. S. Emtage.** 1979. Complete nucleotide sequences of an influenza virus haemagglutinin gene from cloned DNA. Nature (London) **282:**471–477.
28. **Queen, C. L., and L. J. Korn.** 1980. Computer analysis of nucleic acids and proteins. Methods Enzymol. **65:**595–609.
29. **Robertson, J. S.** 1979. 5′ and 3′ terminal nucleotide sequences of the genome segments of influenza virus. Nucleic Acids Res. **6:**3745–3757.
30. **Robertson, J. S., M. Schubert, and R. A. Lazzarini.** 1981. Polyadenylation sites for influenza virus mRNA. J. Virol. **38:**157–163.
31. **Rose, J. K., and C. J. Gallione.** 1981. Nucleotide sequences of the mRNA's encoding the vesicular stomatitis virus G and M proteins determined from cDNA clones containing the complete coding regions. J. Virol. **39:**519–528.
32. **Satake, M., N. Elango, and S. Venkatesan.** 1984. Sequence analysis of the respiratory syncytial virus phosphoprotein gene. J. Virol. **52:**991–994.
33. **Satake, M., and S. Venkatesan.** 1984. Nucleotide sequence of the gene encoding respiratory syncytial virus matrix protein. J. Virol. **50:**92–99.
34. **Shioda, T., Y. Hidaka, T. Kanda, H. Shibuta, A. Nomoto, and K. Iwasaki.** 1983. Sequence of 3,687 nucleotides from the 3′ end of Sendai virus genome RNA and the predicted amino acid sequences of viral NP, P, and C proteins. Nucleic Acids Res. **11:**7317–7330.
35. **Swartz, M. N., T. A. Trautner, and A. Kornberg.** 1962. Enzymatic synthesis of DNA: further studies on nearest neighbor base sequences in deoxyribonucleic acids. J. Biol. Chem. **237:**1961–1967.
36. **Thayer, R. E.** 1979. An improved method for detecting foreign DNA in plasmids of *Escherichia coli*. Anal. Bioch. **98:**60–63.

37. **Thomas, P. S.** 1980. Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose. Proc. Natl. Acad. Sci. U.S.A. **77:**5201–5205.

38. **Venkatesan, S., N. Elango, and R. M. Chanock.** 1983. Construction and characterization of cDNA clones for four respiratory syncytial viral genes. Proc. Natl. Acad. Sci. U.S.A. **80:**1280–1284.

39. **Venkatesan, S., N. Elango, M. Satake, E. Camargo, and R. M. Chanock.** 1984. Organization and expression of respiratory syncytial virus, p. 37–44. *In* R. M. Chanock and R. A. Lerner (ed.), Modern approaches to vaccines: molecular and chemical basis of virus virulence and immunogenicity. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

40. **Wilbur, W. J., and D. J. Lipman.** 1983. Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. U.S.A. **80:**726–730.