JOURNAL OF VIROLOGY, Oct. 1991, p. 5425–5434
0022-538X/91/105425-10$02.00/0

Vol. 65, No. 10

# Genetic Diversity of the Attachment Protein of Subgroup B Respiratory Syncytial Viruses

WAYNE M. SULLENDER,[1,2]* MAURICE A. MUFSON,[3] LARRY J. ANDERSON,[4] AND GAIL W. WERTZ[2]

*Department of Pediatrics[1] and Department of Microbiology,[2] University of Alabama School of Medicine, Birmingham, Alabama 35294-0011; Department of Medicine, Marshall University School of Medicine, and Veterans Administration Medical Center, Huntington, West Virginia 25701[3]; and Respiratory and Enterovirus Branch, Division of Viral Rickettsial Diseases, Centers for Disease Control, Atlanta, Georgia 30333[4]*

Received 15 May 1991/Accepted 4 July 1991

**Respiratory syncytial (RS) virus causes repeated infections throughout life. Between the two main antigenic subgroups of RS virus, there is antigenic variation in the attachment protein G. The antigenic differences between the subgroups appear to play a role in allowing repeated infections to occur. Antigenic differences also occur within subgroups; however, neither the extent of these differences nor their contributions to repeat infections are known. We report a molecular analysis of the extent of diversity within the subgroup B RS virus attachment protein genes of viruses isolated from children over a 30-year period. Amino acid sequence differences as high as 12% were observed in the ectodomains of the G proteins among the isolates, whereas the cytoplasmic and transmembrane domains were highly conserved. The changes in the G-protein ectodomain were localized to two areas on either side of a highly conserved region surrounding four cysteine residues. Strikingly, single-amino-acid coding changes generated by substitution mutations were not the only means by which change occurred. Changes also occurred by (i) substitutions that changed the available termination codons, resulting in proteins of various lengths, and (ii) a mutation introduced by a single nucleotide deletion and subsequent nucleotide insertion, which caused a shift in the open reading frame of the protein in comparison to the other G genes analyzed. Fifty-one percent of the G-gene nucleotide changes observed among the isolates resulted in amino acid coding changes in the G protein, indicating a selective pressure for change. Maximum-parsimony analysis demonstrated that distinct evolutionary lineages existed. These data show that sequence diversity exists among the G proteins within the subgroup B RS viruses, and this diversity may be important in the immunobiology of the RS viruses.**

Respiratory syncytial (RS) virus, a pneumovirus in the family *Paramyxoviridae*, is the most frequently isolated agent in acute respiratory tract infections of children (27, 36). Efforts to develop a safe and effective vaccine have been unsuccessful, and a Formalin-inactivated vaccine tested in the 1960s not only failed to protect but resulted in exacerbated disease among some recipients (25). Several aspects of infections with the RS viruses highlight their unusual and poorly understood biology. The majority of severe RS virus infections occur in children in the first 6 months of life, and infants may be infected even though they possess transplacentally acquired maternal antibody against RS virus. The presence of higher titers of maternal antibody, however, is correlated with less severe disease (15). Additionally, reinfections with RS virus occur throughout life, and these infections occur in the presence of pre-existing antibody (19). It is not known whether antigenic diversity among the viruses or inadequacies of the immune response or both allow reinfections to occur.

The attachment protein G is one of the most unusual features of RS virus (46). The G protein, one of two major transmembrane glycoproteins of RS virus, lacks both the hemagglutinin and neuraminidase activities found in other paramyxovirus attachment proteins. Sequence analysis of the G gene and its deduced amino acid sequence show no homology at the nucleotide or amino acid level with other known RNA virus genes or gene products. Instead, the amino acid sequence reveals a protein containing 30% serine and threonine and 10% proline. Consistent with the high content of serines and threonines, which are potential attachment sites for O-linked sugars, the G protein is estimated to have approximately 50% of its molecular weight contributed by extensive O-linked carbohydrates and some N-linked carbohydrates (47). Thus, the 298-amino-acid, 32,587-$M_r$ protein encoded by the G gene is modified by carbohydrate addition to obtain the mature, highly glycosylated 90,000-$M_r$ form of the protein. The high serine, threonine, and proline contents of the G protein give it an overall resemblance to a class of proteins known as mucins (46).

There are two major antigenic subgroups (A and B) of the RS viruses (4, 30). Between the two subgroups the greatest antigenic and sequence differences are found in the attachment protein G (23). Children initially infected with a subgroup A RS virus appear more likely to experience reinfection with a subgroup B virus (29). Thus, the antigenic differences between subgroups are important in understanding the immune responses to infections and reinfections with the RS viruses. The response elicited by immunization with recombinant vaccinia viruses expressing either the F or G protein of RS virus is capable of restricting viral replication in the lungs of immunized animals. For the G protein, this protection is subgroup specific (24, 39, 41). Additionally, antigenic differences have been found within subgroups. The

* Corresponding author.

TABLE 1. Date and place of isolation and antigenic characterization

| Virus strain | Isolation | | MAb reactivity[a] | | Reference[b] |
|---|---|---|---|---|---|
| | Yr | Place | CDC | WV | |
| 8/60 | 1960 | Sweden | B/2 | B1 | 11 |
| CH18537 | 1962 | District of Columbia | B/3 | B1 | 8 |
| 9320 | 1977 | Massachusetts | B/1 | B1 | 21 |
| WV4843 | 1980–1981 | West Virginia | B/2 | B2 | 32 |
| WV10010 | 1983 | West Virginia | B/2 | B1 | 29 |
| WV15291 | 1985 | West Virginia | B/3 | B1 | 29 |
| NM1355 | 1989 | New Mexico | B/2 | B1 | 40 |

[a] Monoclonal antibody (MAb) reactivity was determined at the Centers for Disease Control (CDC) and Veterans Administration Medical Center, Huntington, W.V. (WV), as previously described (2, 3). These two analyses used different antibodies and classification schemes and thus identify different antigenic variants.

[b] Original citations for the viruses indicated.

G proteins of the subgroup B RS viruses are heterogenous in their reactivity with monoclonal antibodies. Akerlind et al. (2) have identified two groups of antigenic variants designated B1 and B2, and Anderson et al. (3) have described four groups of antigenic variants within the subgroup B RS viruses.

The original sequence analyses of the G gene of RS virus were carried out on prototype viruses isolated approximately 30 years ago. Comparison of two subgroup A and two subgroup B RS virus isolates showed limited sequence variation of the G gene within a subgroup (23, 41). We have obtained viruses isolated from children over the past 30 years and carried out a nucleotide sequence analysis of cDNAs of the G-gene mRNAs of these isolates. In this report we compare the G nucleotide and deduced amino acid sequences of viruses from 1960, 1962, 1977, 1980-1981, 1983, 1985, and 1989, and we define the extent and location of changes. This analysis includes a pair of subgroup B RS virus isolates obtained from sequential infections of a single child. The data presented show significant amino acid sequence variation in the ectodomain of the G protein; this variation was localized to two areas of the ectodomain. This work demonstrates that in natural isolates from children G-protein variation has occurred and that the changes resulted not only from substitution point mutations but also from frameshift mutations due to nucleotide insertions and deletions. Furthermore, the work described here underlines the biologic significance of the recent finding by Garcia-Barreno et al. (14) that a novel mechanism involving frameshift mutations generated monoclonal antibody-resistant mutants of RS virus G protein in vitro.

## MATERIALS AND METHODS

**Cells and viruses.** The subgroup B RS virus isolates 8/60, CH18537, 9320, and NM1355 and their growth in HEp-2 cells have been described previously (8, 11, 21, 40). Additional isolates used in this investigation were WV4843 (28, 32), WV10010, and WV15291. The last two isolates were obtained from a child infected at 14 months of age (WV10010) and again at 36 months of age (WV15291) (29). The year and place of isolation of each virus are shown in Table 1. The subgroup B classification of each virus was determined by reactivity with two different panels of monoclonal antibodies (4, 30).

**Oligonucleotides.** Synthetic oligonucleotides used as primers in the polymerase chain reaction included G32b (ACC CGGGGATCCGCAACCATGTCCAAACACAAG), with bases 13 through 33 corresponding to bases 10 through 30 in the 8/60 G nucleotide sequence (41) and the 12 bases at the 5′ end providing *Sma*I and *Bam*HI restriction sites. G922b (AC CCGGGGATCCGAATAACTAAGCATGTGACTAGG) was complementary (from positions 13 through 35) to bases 921 through 900 in the 8/60 G nucleotide sequence, with bases 1 to 12 providing restriction sites as for G32b. pT-Bam was an oligo-d(T)-based primer [$CGGGATCC(T)_{20}$] with a *Bam*HI restriction endonuclease site added to the 5′ end (5). Internal sequencing primers were designed based on the 8/60 G-nucleotide sequence, and commercially available primers were used for sequencing from plasmid DNA.

**cDNA synthesis and amplification.** Viral mRNA was prepared from infected-cell lysates by extraction with hot phenol (34) and used as a template for the synthesis of cDNA with either G922b or pT-Bam as a primer. Synthesis was performed at 37°C for 45 min with Moloney murine leukemia virus reverse transcriptase (13). The first strand of cDNA was amplified by using the polymerase chain reaction (PCR) (35). Either G922b or pT-Bam was used as a 3′ primer, and the 5′ primer was G32b. Thermal cycling was performed using 94°C for 1 min, 64°C for 1 min, and 72°C for 1 min for 35 cycles, 72°C for 7 min, and a 4°C holding temperature. Analysis by agarose gel electrophoresis showed products of the expected size (0.9 kb) (data not shown). If another amplification was to follow, the original products were purified by using Centricon 100 microconcentrators (Amicon, Beverly, Mass.) or Chromaspin 400 columns (Clontech, Palo Alto, Calif.). For direct sequencing, the double-stranded PCR product was amplified asymmetrically by using a program of 94°C for 1 min, 60°C for 1 min, and 72°C for 1 min for 20 cycles, 72°C for 7 min, and a 4°C holding temperature (12). The asymmetric products were purified as described above and used directly as the template for sequence determination.

**Cloning.** DNA products from the PCR were digested with *Bam*HI for full-length clone preparation or with restriction endonucleases that cleaved the DNA internally to clone smaller fragments. The plasmid vector pGEM-3Z (Promega, Madison, Wis.) was cleaved with the appropriate restriction endonucleases, treated with calf intestinal alkaline phosphatase, ligated with the insert pieces, and used to transform DH5-alpha competent cells. DNA prepared by alkaline lysis (49) was analyzed by restriction endonuclease digestion to identify the clones containing cDNA inserts.

**Nucleotide sequence analysis.** Dideoxynucleotide chain termination sequencing was performed by using as templates denatured plasmid DNA (18) and the products of the asymmetric PCR described above. A modified T7 polymerase (43) and a commercial kit (Sequenase 2.0 DNA sequencing kit; U.S. Biochemical) were used. DNA and deduced amino acid sequence analyses were performed with a VAX computer and the University of Wisconsin Genetics Computer Group programs (10). Phylogenetic analysis was performed by using maximum parsimony with the PAUP software version 3.0 (42).

**Nucleotide sequence accession numbers.** The GenBank accession numbers for the viral G gene nucleotide sequences described here are M73544 for 9320, M73540 for WV4843, M73541 for WV10010, M73542 for WV15291, and M73543 for NM1355.

## RESULTS

**Viruses and monoclonal antibody reactivity.** We obtained subgroup B RS viruses isolated between 1960 and 1989 to carry out a sequence analysis of their G genes. The viruses were isolated in two countries and four regions in the United States (Table 1). Two of the viruses, WV10010 and WV15291, were isolated from the same child during two sequential infections.

The viruses were characterized initially as to subgroup by testing their reactivity with monoclonal antibodies (data not shown). These evaluations were performed in two laboratories (Centers for Disease Control, Atlanta, Ga., and Veterans Administration Medical Center, Huntington, W.V.) with two different panels of monoclonal antibodies. Additional differentiation was made within the B subgroup (Table 1) as previously described (2, 3). The differences between the results from the two laboratories reflected the different monoclonal antibodies and nomenclatures used. We noted two patterns of reactivity of the West Virginia antibodies at the epitope G2 on the subgroup B RS virus attachment protein (2). With the Centers for Disease Control monoclonal antibodies we distinguished three different patterns of reactivity among the viruses tested here by the reactivity of the viruses with monoclonal antibodies recognizing epitopes G12, G5a, and G4 (3). Taken together, the data from monoclonal antibody testing allowed the discrimination of five different patterns of reactivity among the seven viruses (Table 1).

**Nucleotide sequence analysis.** The nucleotide sequences determined from the cDNAs of the mRNAs of the G genes of 9320, WV4843, WV10010, WV15291, and NM1355 are shown in Fig. 1 and compared with those of the subgroup B viruses 8/60 (41) and CH18537 (23). The sequences reported here were determined by direct sequencing of asymmetric PCR products or by the sequencing of two to four cDNA clones; the consensus from three or more cDNA clones was used when there were sequence differences between clones. The 5′ 30 bases of the nucleotide sequence for the 9320, WV4843, WV10010, WV15291, and NM1355 G mRNAs and the 3′ 24 bases for 9320 G mRNA were not defined, because this region corresponded to the location of the primers used for the PCR. The 5′ end of the mRNA has shown little variation and no changes in length or apparent initiation codons among previously analyzed subgroup A or B human RS virus or bovine RS virus G mRNA sequences (23, 26). Therefore we assumed that the 5′ termini are the same length and contain the same available initiation codons as those of the viruses analyzed here. With the exception of the NM1355 G mRNA sequence, all of the nucleotide sequences were the same length: 921 nucleotides. The 9320 mRNA nucleotide sequence was not determined after the termination codon, and a specific nucleotide length was thus not defined. The G mRNA of NM1355 had a three-nucleotide insertion at nucleotides 709 through 711. A three-nucleotide gap after nucleotide 708 was inserted in the 8/60, CH18537, 9320, WV4843, and WV15291 G mRNA sequences to maintain alignment relative to NM1355 (Fig. 1 and 2). For WV10010, a single-nucleotide gap at nucleotide 690 and a two-nucleotide gap after nucleotide 708 were used to maintain alignment. The three viruses from the 1960s and 1970s (8/60, CH18537, 9320) used the same termination codon (UAA) at positions 892 to 894 (Fig. 1 and 2B). The virus isolated in 1989 in New Mexico (NM1355) had a nucleotide change at position 892 and instead used a termination codon (UAG) located at positions 901 to 903 (Fig. 2B). The three viruses isolated in West Virginia in 1980 to 1981, 1983, and 1985 (WV4843, WV10010, WV15291) had nucleotide changes at positions 892, 901, and 903, respectively, and used a termination codon (UAG) at positions 913 through 915. The G gene end sequence (3′ UCAAUAAG, viral RNA sense) for all of these viruses was as previously described (Fig. 2B) (24).

**Amino acid sequence analysis.** The deduced amino acid sequences are shown in Fig. 3. Although the first 30 nucleotides were not determined, the major open reading frame used by all of these recent isolates is compatible with initiation at the same AUG used by the G protein of other human and bovine RS viruses (23, 26). Due to the use of different termination codons and the three-nucleotide insertion in the NM1355 G nucleotide sequence, the predicted lengths in amino acid residues of the G proteins varied. Thus, the G proteins of the viruses isolated in the 1960s and 1970s were 292 amino acids long, the G proteins of the viruses isolated from 1980 to 1985 were 299 amino acids long, and the G protein of the virus isolated in 1989 was 296 amino acids long.

The greatest numbers of amino acid changes were observed in the extracellular domain of G, clustered in two areas separated by a conserved region (double-underlined region in Fig. 3). The four cysteine residues between residues 173 to 186 are conserved in all previously described human and bovine RS virus G-protein sequences (26). Between A and B subgroup viruses a conserved region of 13 amino acids from residues 164 to 176 has been described (23). Among the subgroup B virus G-protein sequences described here, a highly conserved region was observed that extends from 20 amino acids before and 35 amino acids after the cysteine residues (amino acids 153 to 221, inclusive) (Fig. 3).

There were no changes observed in the transmembrane domain, and only two changes were observed in the cytoplasmic tail (with the five N-terminal residues not determined due to the location of the PCR primers). The changes in the ectodomain were located in two groups on either side of the highly conserved region (residues 153 through 221) surrounding the cysteines (Fig. 3B). Several unusual features were observed within the variable region beginning at amino acid 222. In the 12 amino acids from residues 222 to 233, each G sequence differed from the 8/60 sequence, ranging from one change for the CH18537 G-gene sequence to nine differences for the WV10010 G-gene sequence (Fig. 3A). The WV10010 G nucleotide and amino acid sequences were particularly notable. A change in reading frame was observed in comparison to the other G-protein sequences as a result of the deletion of an adenosine from the cluster of adenosines found from nucleotides 684 to 690 in the other isolates (Fig. 1 and 2). This change in reading frame continued until after nucleotide 710, where the addition of an extra adenosine residue, as compared with the other G-gene sequences, resulted in a shift back into the original reading frame. In the region of the apparent shift in reading frame, the WV10010 G protein had seven consecutive amino acids (residues 226 through 232) that differed from those in the other G-protein sequences (Fig. 3A). Immediately after this frameshift region in the WV10010 sequence, a noteworthy sequence change was observed for another virus, NM1355, which gained three nucleotides (after position 708 in the nucleotide sequence, Fig. 1 and 2), resulting in the addition of a threonine residue at position 233 in the amino acid sequence (Fig. 3A).

**Glycosylation sites.** The G protein of RS virus is modified

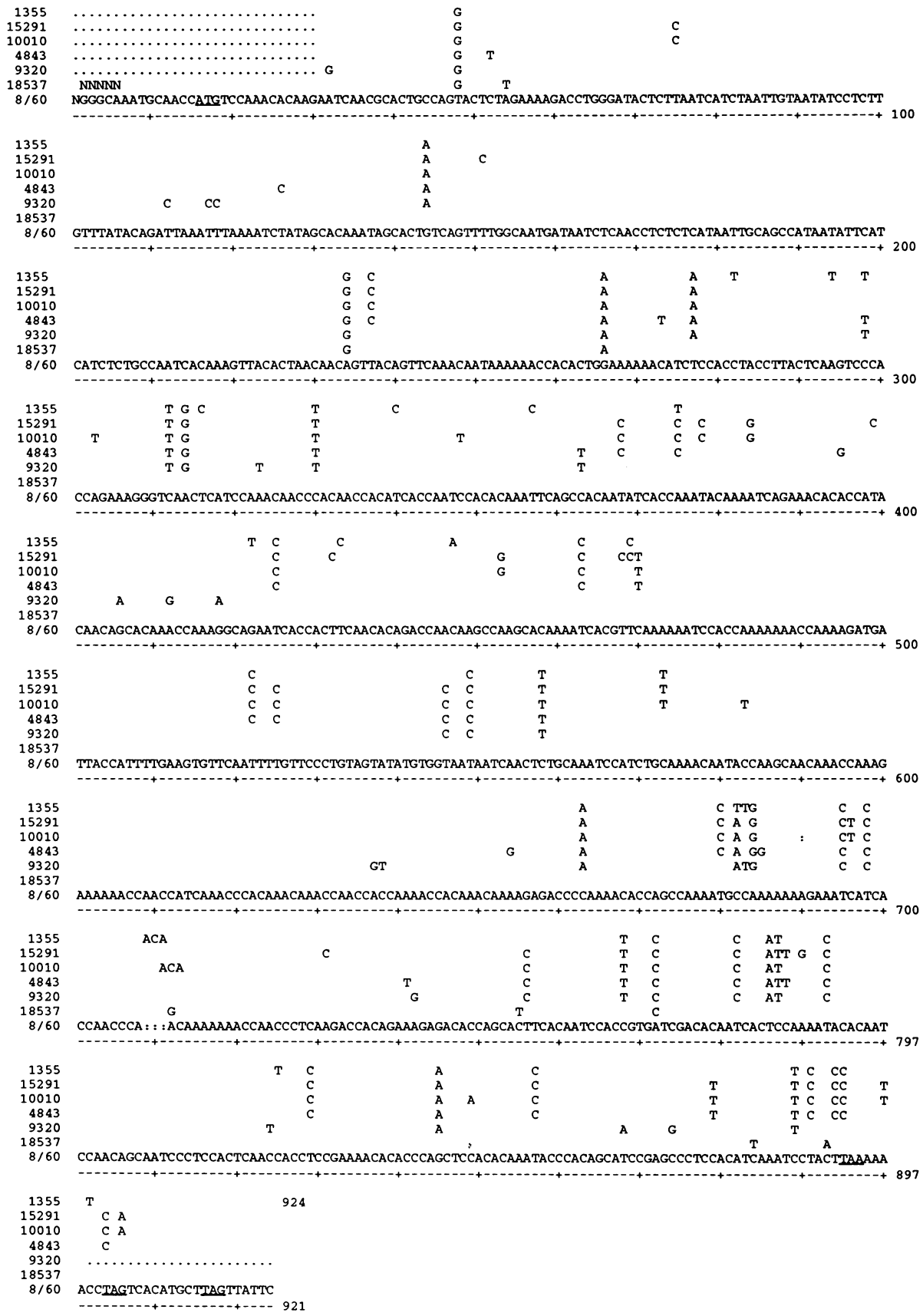
1355
15291
10010
4843
9320
18537
8/60
NNNNN
NGGGCAAATGCAACCATGTCCAAACACAAGAATCAACGCACTGCCAGTACTCTAGAAAAAGACCTGGGATACTCTTAATCATCTAATTGTAATATCCTCTT
100
GTTTATACAGATTAAATTTAAAATCTATAGCACAAATAGCACTGTCAGTTTTGGCAATGATAATCTCAACCTCTCTCATAATTGCAGCCATAATATTCAT
200
CATCTCTGCCAATCACAAAGTTACACTAACAACAGTTACAGTTCAAACAATAAAAAACCACACTGGAAAAAACATCTCCACCTACCTTACTCAAGTCCCA
300
CCAGAAAGGGTCAACTCATCCAAACAACCCACAACCACATCACCAATCCACACAAATTCAGCCACAATATCACCAAATACAAAATCAGAAACACACCATA
400
CAACAGCACAAACCAAAGGCAGAATCACCACTTCAACACAGACCAACAAGCCAAGCACAAAATCACGTTCAAAAAATCCACCAAAAAAACCAAAAGATGA
500
TTACCATTTTGAAGTGTTCAATTTTGTTCCCTGTAGTATATGTGGTAATAATCAACTCTGCAAATCCATCTGCAAAACAATACCAAGCAACAAACCAAAG
600
AAAAAACCAACCATCAAACCCACAAACAAACCAACCACCAAAACCACAAACAAAAGAGACCCCAAAACACCAGCCAAAATGCCAAAAAAAGAAATCATCA
700
CCAACCCA:::ACAAAAAAACCAACCCTCAAGACCACAGAAAGAGACACCAGCACTTCACAATCCACCGTGATCGACACAATCACTCCAAAATACACAAT
797
CCAACAGCAATCCCTCCACTCAACCACCTCCGAAAACACACCCAGCTCCACACAAATACCCACAGCATCCGAGCCCTCCACATCAAATCCTACTTAAAAA
897
924
ACCTAGTCACATGCTTAGTTATTC
921

FIG. 1. Nucleotide sequence comparisons among the attachment protein G mRNAs of the subgroup B RS viruses as determined from cDNA sequence analysis. The sequence of the 8/60 G cDNA is shown; for the other viruses only differences relative to 8/60 are indicated. Gaps (:) introduced to maintain alignment and sequences not determined due to the location of the amplification primers (.) are indicated. Initiation and termination codons are underlined. The 8/60 and CH18537 G sequences were previously described (23, 41).

by the addition of both N-linked and O-linked oligosaccharide chains (46). Based on the comparisons of the amino acid sequences, the number of sites for potential N-linked glycosylation (Asn-X-Ser/Thr) varied (Fig. 3). All seven amino acid sequences shared N-linked glycosylation sites at amino acids 81 and 86, with the site at residue 86 conserved even though a change occurred from serine in the 8/60 and CH18537 sequences to threonine in the amino acid sequences of the other isolates. A third site in the 8/60 and CH18537 G amino acid sequences at residue 100 was not present in the sequences of the more recent isolates. However, the other five isolates gained sites for potential N-linked sugar addition at residues 276 and 290, for a total of four potential sites in the 9320, WV4843, WV15291, WV10010, and NM1355 G proteins compared with three sites in the 8/60 and CH18537 G proteins. All of these amino acid sequences had elevated contents of serine and threonine (range, 28 to 32%), which are potential O-linked glycosylation sites, and elevated proline contents of 7 to 9% as previously described for human and bovine RS virus G-protein sequences (26, 46).

**Sequence comparisons.** The numbers of G nucleotide and amino acid differences between the G genes and proteins of pairs of viruses (Table 2) were calculated from the data shown in Fig. 1 and 3. Two viruses from the early 1960s (8/60 and CH18537), although isolated on different continents, had G-gene and -protein sequences that were quite similar, with nine nucleotide differences and seven amino acid differences. When the 8/60 G protein was compared with the G proteins of viruses from the 1970s and 1980s, over three times as many amino acid differences (range, 22 to 27) were found as between the 8/60 and CH18537 G proteins. The G protein of the virus from 1977 (9320) had a similar number of amino acid differences (range, 22 to 27) compared with the G proteins of viruses from the 1960s and the 1980s. The G proteins of the four viruses from the 1980s (WV4843, WV10010, WV15291, NM1355) were more similar to one another, with the number of amino acid differences ranging from 7 to 21. Among the G proteins of these four viruses, the three isolated in West Virginia (WV4843, WV10010, WV15291) had fewer differences from one another, whereas the G protein of the isolate from New Mexico (NM1355) had a higher number of differences from the West Virginia viruses.

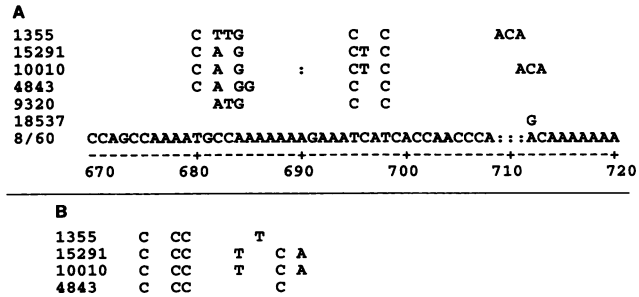


FIG. 2. (A) Region of glycoprotein G-gene nucleotide sequence in which insertions and deletions were observed. The 8/60 G cDNA sequence from bases 670 to 717 is shown; for the other viruses only nucleotides that differ from 8/60 are shown. Gaps (:) were introduced to maintain alignment. (B) Termination codons used by the attachment proteins of the subgroup B RS viruses. The 8/60 nucleotide sequence is shown from bases 887 to 921, and the other sequences indicate differences relative to 8/60. For 9320, the 3' sequence not determined due to the location of the amplification primer is marked (.). The three termination codons are underlined in the 8/60 sequence. 8/60, CH18537, and 9320 use the first termination codon, NM1355 uses the second termination codon, and WV4843, WV10010, and WV15291 use the third termination codon.

The percent differences of the amino acid sequences in the ectodomain of the G protein were also compared (Table 2). These differences ranged from 3% (between 8/60 and CH18537 and between WV4843 and WV15291) to 12% (between 9320 and WV10010 and between NM1355 and 8/60 or CH18537). The amino acid differences in the ectodomain were greater than the differences for the entire protein by 1 to 3%; there were no changes in the transmembrane domain and very few changes in the cytoplasmic domain.

**Nucleotide changes resulting in amino acid changes.** The percentage of nucleotide differences that resulted in amino acid differences was calculated (Table 2). These values ranged from 29% (WV15291 versus WV4843) to 78% (8/60 versus CH18537). A mean (±standard error) of 51% (±2%) of the nucleotide differences resulted in coding changes. This value was greater than the 24% value that would be expected from random nucleotide sequence changes and was similar to the 50% nucleotide differences resulting in amino acid changes described for the HA and NA proteins of strains of influenza A virus (1). For the influenza A viruses, the 50% nucleotide differences resulting in amino acid changes provide evidence for a selective effect of the host immune system for change in the viral proteins.

**Evolutionary relationships.** Phylogenetic analysis of the nucleotide and deduced amino acid sequences was performed by the maximum-parsimony method, and evolutionary trees were constructed (Fig. 4). Multiple evolutionary lineages were observed. The topologies of the nucleotide and amino acid trees were similar even though the overall tree and branch lengths varied. However, the WV10010 branch length in the amino acid tree was greater than that in the nucleotide tree, probably due to the frameshift-induced coding changes. The similarities in the topologies of the two trees were in agreement with the high percentage of nucleotide changes that resulted in amino acid changes, as described above. If a lower proportion of the nucleotide changes resulted in amino acid changes, then a collapse of branch length might be seen in the amino acid tree as compared with that in the nucleotide sequence tree (16).

**Sequential isolates.** The viruses WV10010 and WV15291 were isolated from a single child during respiratory tract infections that occurred 2 years apart. The G sequences had 14 differences at the nucleotide level and 11 differences at the

A.

```
1355  .....     R
15291 .....     R
10010 .....     R
4843  .....     R
9320  .....S    R
18537           R
8/60  MSKHKNQRTASTLEKTWDTLNHLIVISSCLYRLNLKSIAQIALSVLAMIISTSLIIAAIIFIISANHKVTLTTVTVQTIK
      ---------+---------+---------+---------+---------+---------+---------+---------+ 80

1355     E   T    H S    SP       P                          IT  P   K     P P
15291    E   T           S                  T                 T           P L
10010    E   T           S                  T                 T           P L
4843     E   T      S    S                  T        R        T           P L
9320     E   T      S    S  I                        T A S
18537    E
8/60  NHTGKNISTYLTQVPPERVNSSKQPTTTSPIHTNSATISPNTKSETHHTTAQTKGRITTSTQTNKPSTKSRSKNPPKKPK
      Δ----Δ---+---------Δ---------+---------+---------+---------+---------+---------+ 160

1355                                                              TL   TT    T
15291                                                             TT   TT          T
10010                              S                              TT  KLPPTQQ
4843                                                              TTE  TT
9320                                                    V          M   TT
18537                                                                      A
8/60  DDYHFEVFNFVPCSICGNNQLCKSICKTIPSNKPKKKPTIKPTNKPTTKTTNKRDPKTPAKMPKKEIITNPT:KKPTLKT
      ---------+--•--•---+-•---•---+---------+---------+---------+---------+---------+ 239

1355              L  T S H          P    N   T          S QKI     296
15291             L  T LEH          P    N   T          S QNTQSHA
10010             L  T S H          P    N   T          S QNTQSHA
4843   D          L  T L H          P    N   T          S QKTQSHA 299
9320    G         L  T S H        I      N       T A    S
18537       I     L                                  L  N
8/60  TERDTSTSQSTVIDTITPKYTIQQQSLHSTTSENTPSSTQIPTASEPSTSNPT::::::: 292
      ---------+---------+---------+------Δ--+---------+Δ--------+
```

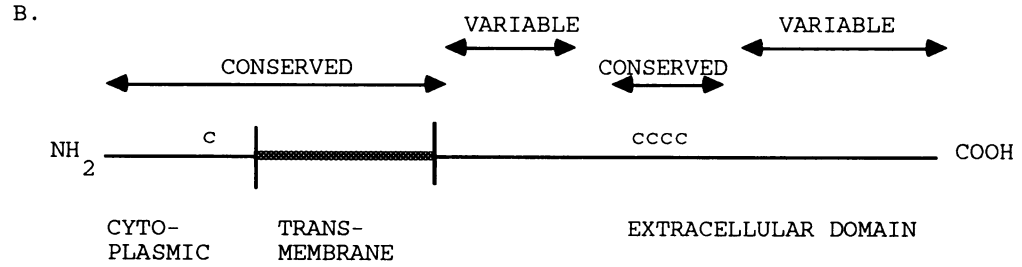


FIG. 3. (A) Comparison of the deduced amino acid sequences among the attachment proteins of the subgroup B respiratory syncytial viruses. The sequence of the 8/60 G protein is shown; for the other viruses only residues that differ from the 8/60 sequence are indicated. Gaps (:) were introduced to maintain alignment, cysteines in the ectodomain (●) are indicated, as are sites for potential N-linked glycosylation (Δ). The transmembrane domain is underlined with a single line, the conserved region from residues 153 to 221 is underlined with a double line, and the sequence regions not determined due to the location of the amplification primers (.) are indicated. The 8/60 and CH18537 sequences were previously described (23, 41). (B) Linear representation of the glycoprotein G of the RS viruses. The postulated domains of the protein and the cysteine residues (C) are as shown. The conserved regions of the cytoplasmic tail and transmembrane domain, the region surrounding the four cysteine residues in the ectodomain, and the more variable regions in the ectodomain are indicated.

amino acid level, providing a 96% total amino acid identity. An extremely high percentage of nucleotide changes (69%) thus result in coding changes (Table 2). Of the 11 amino acid changes, all are in the extracellular domain, with 7 consecutive changes (positions 226 through 232) in the frameshift region described above. These two sequences have no amino acid differences in the 40 C-terminal amino acids. Antigenic analysis with monoclonal antibodies (Table 1) revealed a difference between these two isolates at one epitope.

TABLE 2. Comparisons among G genes and proteins of the subgroup B RS viruses[a]

| Strain | No. of nucleotide/amino acid differences (% amino acid differences in the ectodomain) from strain: | | | | | | |
|---|---|---|---|---|---|---|---|
| | 8/60 | 18537 | 9320 | 4843 | 10010 | 15291 | 1355 |
| 8/60 | | 9/7 (3) | 42/24 (10) | 51/24 (10) | 57/26 (11) | 57/22 (9) | 51/27 (12) |
| 18537 | **78** | | 43/25 (11) | 52/25 (11) | 57/26 (11) | 58/23 (10) | 52/28 (12) |
| 9320 | **57** | **58** | | 38/22 (9) | 47/27 (12) | 47/22 (9) | 43/24 (10) |
| 4843 | **47** | **48** | **58** | | 26/14 (6) | 24/7 (3) | 34/15 (7) |
| 10010 | **46** | **46** | **57** | **54** | | 16/11 (5) | 38/21 (9) |
| 15291 | **39** | **40** | **47** | **29** | **69** | | 38/16 (7) |
| 1355 | **53** | **54** | **56** | **44** | **55** | **42** | |

[a] The data in lightface type indicate the total nucleotide and amino acid differences for the G genes or proteins and, within parentheses, the percent amino acid differences in the ectodomain. These values were counted from the alignments shown in Fig. 1 and 2; gaps at the beginnings and ends of the sequences were not counted as differences, gaps within sequences were counted as differences. The number of amino acid differences divided by the number of nucleotide differences was used for the calculation of the percent nucleotide differences resulting in amino acid coding changes, indicated by numbers in boldface type.

## DISCUSSION

In this report we describe the analysis of the sequences of the G genes of subgroup B RS viruses isolated over the past 30 years. The pertinent observations made from this data were as follows. (i) Significant variation occurred among the G protein sequences within an RS virus subgroup. (ii) Fifty-one percent of the nucleotide changes resulted in amino acid coding changes for the G proteins among viruses, and most of these amino acid changes occurred in the ectodomain of the protein, suggesting a selective effect of the host immune response. (iii) Multiple evolutionary lineages were observed. (iv) Nucleotide substitutions caused both point mutations and differences in protein length due to termination codon changes. (v) A frameshift mutation was observed in one isolate.

These data reveal greater sequence heterogeneity within a single RS virus subgroup than has been previously reported and greater differences than would have been predicted by reaction patterns against two panels of monoclonal antibodies (Table 1). The sequence data and either panel of monoclonal antibodies gave different pictures of the similarities and differences among the isolates. Eventual identification of the sequences encoding the G-protein epitopes should help resolve the differences between the two monoclonal antibody reactivity classification schemes used here (2, 3). Comparisons between two subgroup B RS virus G proteins (8/60 and CH18537, isolated in 1960 and 1962, respectively) reveal 2% amino acid differences (41), whereas two subgroup A RS viruses (Long and A2), isolated in 1956 and 1961, respectively, have 7% amino acid differences between their G proteins (23). In this investigation we found that, among the subgroup B RS viruses, amino acid differences of 9% occurred between the full-length proteins, whereas analysis of the G-protein ectodomains showed amino acid differences as great as 12%. In addition, the mean value of 51% nucleotide changes resulting in amino acid changes was greater than that expected from randomly occurring mutations. For comparison, among influenza A virus drift strains there are amino acid differences of 10% among the HA proteins and about 50% of the nucleotide changes result in amino acid changes (1). These differences among influenza A virus drift strains are thought to be due to the selective effect of antibody pressure and are associated with important changes in antigenicity, allowing reinfections to occur (33). Potentially, the changes observed in the attachment protein G or yet-to-be-described changes in other RS virus proteins may contribute to the occurrence of reinfections with the RS

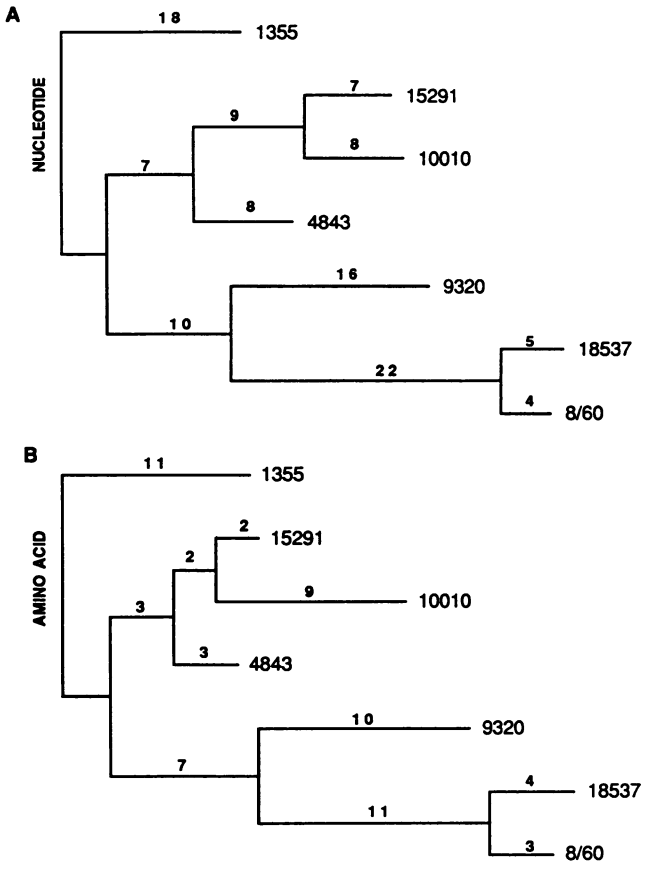


FIG. 4. Evolutionary trees for the subgroup B RS virus attachment protein nucleotide (A) and amino acid (B) sequences determined by maximum-parsimony analysis with PAUP software (42). The analysis was performed by using the alignments shown in Fig. 1 and 3. Gaps were assigned as missing values, except that gaps within the nucleotide or amino acid sequences or at the C termini of the amino acid sequences were counted as one difference and then were assigned as missing values if longer than one residue. The options used included maxtree = 100, MULPARS, branch and bound, and unrooted; trees were rooted by using a default outgroup method. The nucleotide tree length is 114 and the amino acid tree length is 65. The lengths of the horizontal branch lines are shown and are proportional to the minimum number of changes necessary to produce the observed differences. The vertical line lengths are arbitrary, and the vertical placement reflects the chronological order of entry of the sequence data.

viruses and pose an obstacle to the development of broadly protective RS virus vaccines.

Analysis by the method of maximum parsimony of the subgroup B RS virus attachment protein nucleotide and amino acid sequences revealed the existence of distinct evolutionary lineages rather than a single lineage (Fig. 4). Thus, the evolutionary pattern for RS viruses appeared to be more similar to that of the influenza B or C viruses, which have multiple lineages, rather than that of the single evolutionary lineage described for the influenza A viruses (1, 48). However, a clearer picture of the evolutionary lineages of the RS viruses will require nucleotide sequence determination studies of additional isolates, including viruses from the same and different communities within a single epidemic period. Other investigators have examined such viruses by testing their reactivity with monoclonal antibodies and by using RNase A protection assays, demonstrating that within an epidemic period there is heterogeneity among the same subgroup viruses (3, 9, 20, 37, 38). An investigation of both temporally distinct and single-outbreak isolates has also been performed with PCR-amplified products of the SH and N genes (6). The sequence changes that occurred in the SH gene were limited and resulted in less of a change in amino acid identity than was observed in this investigation; however, several evolutionary lineages were described for the subgroup A RS viruses based on differences among N and among SH genes.

The G-gene and -protein sequences of two sequential isolates (WV10010 and WV15291) from a single child were compared. There were 17 nucleotide differences and 11 amino acid differences between these two viruses, resulting in an amino acid identity of 95% for the G-protein ectodomain. A frameshift mutation occurred in the first isolate (WV10010); the deletion of an adenosine resulted in a change of reading frames, and the subsequent insertion of an adenosine caused a shift back into the original reading frame. This frame shift caused coding changes with seven amino acids in the WV10010 G-protein sequence that were completely different from those of all of the other G-protein sequences, including that of WV15291, the second of the sequential isolates. It is interesting to note that monoclonal antibody selection of a subgroup A RS virus, the Long strain, has resulted in mutations occurring by a frameshift mechanism (14). Although none of the changes described among these in vitro variants occurred at the positions observed in this investigation, in both studies insertions and deletions were found at clusters of adenosine residues. The mechanism for such changes remains to be defined; a "slippery polymerase" that shifts on the RNA template during replication has been postulated (14) and may be similar to the stuttering described during polyadenylation of mRNA (17) or to the various forms of mRNA editing in the paramyxoviruses (7, 44, 45). The WV10010 virus had never been subjected to in vitro antibody selection. These findings show that naturally occurring RS viruses have the capacity for extensive amino acid changes resulting from frameshift mutations.

Another mechanism for change among the subgroup B viruses involved the use of different termination codons (Fig. 3). Of the three termination codons present in the 8/60, CH18537, and 9320 sequences, the first and/or second termination codon was not present in the G genes of viruses isolated at later times. The resultant changes in peptide length, in conjunction with differing numbers of sites (three to four) for potential N-linked glycosylation, may explain the differing mobilities reported for G proteins within an individual subgroup (31).

The data presented herein provide strong evidence of a selection for change in the attachment protein of the subgroup B RS viruses. The selective pressure for this change is likely to be the antibody response for the host. Three items support this hypothesis. (i) The greatest number of amino acid changes was in the ectodomain of the G protein, the domain of the protein exposed to host immune surveillance by the antibody. (ii) Over half of the nucleotide differences caused amino acid changes, which is a greater percentage than would result from random mutations. (iii) Changes were introduced not only by substitutions but also by frameshift mutations, in a manner similar to that described previously for monoclonal antibody escape mutants of the RS virus G protein in vitro (14). Thus it appears that there is a positive selection for change in the G protein, as has been described for the HA and NA proteins of drift strains of the influenza A viruses (1). A positive selection indicates that the virus derives some evolutionary advantage from the changes it acquires, suggesting that the avoidance of host antibody recognition benefits the virus, possibly by facilitating repeated infections with the RS viruses.

Previous work has shown that the location of the four cysteines in the ectodomain of the G protein is common to human and bovine RS viruses. Furthermore, a strictly conserved region of 13 amino acids, partly overlapping with the cysteines, is present in human A and B subgroup RS virus G proteins but not in the bovine RS virus G protein (23, 27). In contrast, among the subgroup B virus G proteins described here, a larger region of 69 amino acids surrounding the cysteines was highly conserved with only two amino acid changes. This lends support to the hypothesis that this conserved region may be particularly important in terms of the attachment function or the conformation required for the attachment function of this protein (23).

In summary, we have described an analysis of subgroup B RS virus G proteins. Significant nucleotide and deduced amino acid sequence differences and multiple evolutionary lineages were observed. In addition to substitution changes in the sequences, changes originated by frameshift mutation. We found evidence for a positive selection for change in the G protein by antibody recognition. These changes may result in important immunologic differences between these viruses, and future studies will address these questions.

## ACKNOWLEDGMENTS

Support was received from Public Health Service grants F32 AI107864 (to W.M.S.) and R37 AI2464 and AI20181 (to G.W.W.) from the National Institutes of Health from the World Health Organization Programme on Vaccine Development. Support for maintenance of the Wisconsin-GCG computer programs utilized here was provided by the Center for AIDS Research P30 AI27767. Oligonucleotides were made in the Cancer Center DNA Synthesis Core Facility (supported by Public Health Service grant CA13148 from the National Institutes of Health).

We thank Mattanja Coehoorn, Tanya Kelly, Patrick Dean, and Lirong Sun for assistance with the nucleotide sequence determinations.

## REFERENCES

1. **Air, G. M., A. J. Gibbs, W. Graeme-Laver, and R. G. Webster.** 1990. Evolutionary changes in influenza B are not primarily governed by antibody selection. Proc. Natl. Acad. Sci. USA **87**:3884–3888.
2. **Akerlind, B., E. Norby, C. Orvell, and M. A. Mufson.** 1988. Respiratory syncytial virus: heterogeneity of subgroup B

strains. J. Gen. Virol. **69**:2145–2154.
3. **Anderson, L. J., R. M. Hendry, L. T. Pierik, C. Tsou, and K. Mcintosh.** 1991. Multicenter study of strains of respiratory syncytial virus. J. Infect. Dis. **163**:687–692.
4. **Anderson, L. J., J. C. Hierholzer, C. Tsou, R. M. Hendry, B. F. Fernie, Y. Stone, and K. Mcintosh.** 1985. Antigenic characterization of respiratory syncytial virus strains with monoclonal antibodies. J. Infect. Dis. **157**:626–633.
5. **Berchtold, M. W.** 1989. A simple method for direct cloning and sequencing cDNA by the use of a single specific oligonucleotide and oligo(dT) in a polymerase chain reaction (PCR). Nucleic Acids Res. **17**:453.
6. **Cane, P. A., and C. R. Pringle.** 1991. Respiratory syncytial virus heterogeneity during an epidemic: analysis by limited nucleotide sequencing (SH gene) and restriction mapping (N gene). J. Gen. Virol. **72**:349–357.
7. **Cattaneo, R., K. Kaelin, K. Baczko, and M. A. Billeter.** 1989. Measles virus editing provides an additional cysteine-rich protein. Cell **56**:759–764.
8. **Coates, H. V., L. Kendrick, and R. M. Chanock.** 1963. Antigenic differences between two strains of respiratory syncytial virus. Proc. Soc. Exp. Biol. Med. **112**:958–964.
9. **Cristina, J., J. A. Lopez, C. Albo, B. Garcia-Barreno, J. Garcia, J. A. Melero, and A. Portela.** 1990. Analysis of genetic variability in human respiratory syncytial virus by the RNase A mismatch cleavage method: subtype divergence and heterogeneity. Virology **174**:126–134.
10. **Devereux, J., P. Haberli, and O. Smithies.** 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12**:387–395.
11. **Doggett, J. E., and D. Taylor-Robinson.** 1965. Serological studies with respiratory syncytial virus. Arch. Gesamte Virusforsch. **15**:601–608.
12. **Dorit, R. L., and O. Ohar.** 1990. Direct DNA sequencing of polymerase chain reaction products, p. 15.2.1–15.2.11. *In* F. M. Ausubel (ed.), Current protocols in molecular biology. Greene Publishing Associates, New York.
13. **Ferre, F., and F. Garduno.** 1989. Preparation of crude cell extract suitable for amplification of RNA by the polymerase chain reaction. Nucleic Acids Res. **17**:2141–2142.
14. **Garcia-Barreno, A. Portela, T. Delgado, J. A. Lopez, and J. A. Melero.** 1990. Frame shift mutations of neutralization resistant mutants of human respiratory syncytial virus. EMBO J. **9**:4181–4187.
15. **Glezen, W. P., A. Paredes, J. E. Allison, L. H. Taber, and A. L. Frank.** 1981. Risk of respiratory syncytial virus infection for infants from low-income families in relationship to age, sex, ethnic group and maternal antibody level. J. Pediatr. **92**:708–715.
16. **Gorman, O. T., R. O. Donis, J. Kawaoka, and R. B. Webster.** 1990. Evolution of influenza A virus PB2 genes: implications for evolution of the ribonucleoprotein complex and origin of human influenza A virus. J. Virol. **64**:4893–4902.
17. **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**:3829–3841.
18. **Haltiner, M., T. Kempe, and R. Tjian.** 1985. A novel strategy for constructing clustered point mutations. Nucleic Acids Res. **13**:1015–1025.
19. **Henderson, F. W., A. M. Collier, W. A. Clyde, and F. W. Denny.** 1979. Respiratory syncytial virus infections: reinfections and immunity: a prospective, longitudinal study in young children. N. Engl. J. Med. **300**:530–534.
20. **Hendry, R. M., L. T. Pierik, and K. Mcintosh.** 1989. Prevalence of respiratory syncytial virus over six consecutive outbreaks: 1981–1987. J. Infect. Dis. **160**:185–190.
21. **Hierholzer, J. C., and M. S. Hirsch.** 1979. Croup and pneumonia in human infants associated with a new strain of respiratory syncytial virus. J. Infect. Dis. **140**:826–828.
22. **Johnson, P. R., and P. L. Collins.** 1988. The A and B subgroups of human respiratory syncytial: comparison of intergenic and gene-overlap sequences. J. Gen. Virol. **69**:2901–2906.
23. **Johnson, P. R., M. K. Spriggs, R. A. Olmstead, and P. L. Collins.** 1987. The G glycoprotein of human respiratory syncytial virus of subgroups A and B: extensive sequence divergence between antigenically related proteins. Proc. Natl. Acad. Sci. USA **84**:5625–5629.
24. **Johnson, P. R., Jr., R. A. Olmstead, G. A. Prince, B. R. Murphy, D. W. Alling, E. E. Walsh, and P. L. Collins.** 1987. Antigenic relatedness between glycoproteins of human respiratory syncytial virus subgroups A and B: evaluation of the contributions of F and G glycoproteins to immunity. J. Virol. **61**:3163–3166.
25. **Kim, H. W., J. G. Canchola, C. D. Brandt, G. Pyles, R. M. Chanock, K. Jensen, and R. H. Parrott.** 1969. Respiratory syncytial virus disease in infants despite prior administration of antigenic inactivated vaccine. Am. J. Epidemiol. **89**:422–434.
26. **Lerch, R. A., K. Anderson, and G. W. Wertz.** 1990. Nucleotide sequence analysis and expression from recombinant vectors demonstrate that the attachment protein G of bovine respiratory syncytial virus is distinct from that of human respiratory syncytial virus. J. Virol. **64**:5559–5569.
27. **McIntosh, K., and R. M. Chanock.** 1990. Respiratory syncytial virus, p. 1045–1072. *In* B. N. Fields et al. (ed), Virology, 2nd ed. Raven Press, New York.
28. **Mufson, M. A., B. Akerlind-Stopne, C. Orvel, R. B. Belshe, and E. Norrby.** 1991. A single-season epidemic with respiratory syncytial virus subgroup B2 during 10 epidemic years, 1978 to 1988. J. Clin. Microbiol. **29**:162–165.
29. **Mufson, M. A., R. B. Belshe, C. Orvell, and E. Norrby.** 1987. Subgroup characteristics of respiratory syncytial virus strains recovered from children with two consecutive infections. J. Clin. Microbiol. **25**:1536–1539.
30. **Mufson, M. A., C. Orvell, B. Rafnar, and E. Norrby.** 1985. Two distinct subtypes of human respiratory syncytial virus. J. Gen. Virol. **66**:2111–2124.
31. **Norrby, E., M. A. Mufson, and H. Sheshberadaran.** 1986. Structural differences between subtype A and B strains of respiratory syncytial virus. J. Gen. Virol. **67**:2721–2729.
32. **Orvell, C., E. Norrby, and M. A. Mufson.** 1987. Preparation and characterization of monoclonal antibodies directed against five structural components of human respiratory syncytial virus subgroup B. J. Gen. Virol. **68**:3125–3135.
33. **Palese, P., and J. F. Young.** 1982. Variation of influenza A, B, and C viruses. Science **215**:1468–1473.
34. **Penman, S.** 1969. Preparation of purified nuclei and nucleoli from mammalian cells, p. 35–48. *In* K. Habel and N. P. Salzman (ed.), Fundamental techniques in virology. Academic Press, Inc., New York.
35. **Saiki, R. K., S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, and N. Arnheim.** 1985. Enzymatic amplification of B globulin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science **230**:1350–1354.
36. **Selwyn, B. J.** 1990. The epidemiology of acute respiratory tract infection in young children: comparison of findings from several developing countries. Rev. Infect. Dis. **2**:5870–5888.
37. **Storch, G. A., L. J. Anderson, C. S. Park, C. Tsou, and D. E. Dohner.** 1991. Antigenic and genomic diversity within group A respiratory syncytial virus. J. Infect. Dis. **163**:858–861.
38. **Storch, G. A., C. S. Park, and D. E. Dohner.** 1989. RNA fingerprinting of respiratory syncytial virus using ribonuclease protection. J. Clin. Invest. **83**:1894–1902.
39. **Stott, E. J., G. Taylor, L. A. Ball, K. Anderson, K.-Y. Young, A. M. Q. King, and G. W. Wertz.** 1987. Immune and histopathological responses in animals vaccinated with recombinant vaccinia viruses that express individual genes of human respiratory syncytial virus. J. Virol. **61**:3855–3861.
40. **Sullender, W., L. J. Anderson, K. Anderson, and G. W. Wertz.** 1990. Differentiation of respiratory syncytial virus subgroups with cDNA probes in a nucleic acid hybridization assay. J. Clin. Microbiol. **28**:1683–1687.
41. **Sullender, W., K. Anderson, and G. Wertz.** 1990. The respiratory syncytial virus subgroup B attachment glycoprotein: analysis of sequence, expression from a recombinant vector, and evaluation as an immunogen against homologous and heterolo-

gous subgroup virus change. Virology **178**:195–203.
42. **Swofford, D. L.** 1990. PAUP: phylogenetic analysis using parsimony, version 3.0. Computer program distributed by the Illinois Natural History Survey, Champaign.
43. **Tabor, S., and C. C. Richardson.** 1987. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. Proc. Natl. Acad. Sci. USA **84**:4767–4771.
44. **Thomas, S. M., R. A. Lamb, and R. G. Paterson.** 1988. Two mRNAs that differ by two non-templated nucleotides encode the amino coterminal proteins P and V of the paramyxovirus SV5. Cell **54**:891–902.
45. **Vidal, S., J. Curran, and D. Kolakofsky.** 1990. A stuttering model for paramyxovirus P mRNA editing. EMBO J. **9**:2017–2022.
46. **Wertz, G. W., P. L. Collins, Y. Huang, C. Gruber, S. Levine, and L. A. Ball.** 1985. Nucleotide sequence of the G protein gene of human respiratory syncytial virus reveals an unusual type of viral membrane protein. Proc. Natl. Acad. Sci. USA **82**:4075–4079.
47. **Wertz, G. W., M. Krieger, and L. A. Ball.** 1989. Structure and cell surface maturation of the attachment glycoprotein of human respiratory syncytial virus in a cell line deficient in *O* glycosylation. J. Virol. **63**:4767–4776.
48. **Yamashita, M., M. Krystal, W. M. Fitch, and P. Palese.** 1988. Influenza B virus evolution: co-circulating lineages and comparison of evolutionary pattern with those of influenza A and C viruses. Virology **163**:112–122.
49. **Zhou, C.** 1990. Ten minute miniprep. Biotechniques **8**:172–173.