Identification of variable domains of the attachment (G) protein of subgroup A respiratory syncytial viruses

P. A. Cane,* D. A. Matthews and C. R. Pringle

Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, U.K.

We have previously classified isolates from a respiratory syncytial (RS) virus epidemic into distinct lineages by restriction mapping and nucleotide sequencing of parts of the nucleocapsid protein and small hydrophobic protein genes, which are areas of the genome not considered to be under immunological pressure. This study has now been extended by the determination of the nucleotide sequences of the attachment (G) protein genes of isolates from each subgroup A lineage. Deduced amino acid identities of the G proteins ranged between 80% and 99%, corresponding closely to the previously determined relatedness of the lineages. The amino acid variability was not evenly distributed; in the extracellular part of the protein there was a sharply defined hypervariable domain which was separated from a more extended variable domain by a highly conserved region. Most nucleotide changes in the variable domains were in the first and second positions of the codon triplets. These results suggest that there may be considerable immunological pressure for change in certain areas of the G protein and this may account for the ability of this virus to reinfect individuals repeatedly. The results presented here reflect the pattern of published data comparing prototype strains of the A and B subgroups.

Introduction

Respiratory syncytial (RS) virus commonly causes severe lower respiratory tract illness in babies and young children, together with usually milder upper respiratory tract infections in older children and adults. An unusual feature of this virus is that it can repeatedly reinfect individuals; this may be due either to an inadequate immune response to the virus or to variability of the virus.

RS virus isolates can be divided into two distinct serological subgroups, A and B, on the basis of their reactions with monoclonal antibodies (MAbs) (Anderson et al., 1985; Mufson et al., 1985; Gimenez et al., 1986). With respect to the attachment (G) protein, MAbs have been shown that there is much antigenic diversity within the subgroups (Morgan et al., 1987; Örvell et al., 1987; Storch & Park, 1987; Garcia-Barreno et al., 1989). Sequencing studies have shown that the G protein is the most variable protein between the subgroups, with only 53% amino acid identity between the G proteins of the A and B prototype strains (Johnson et al., 1987). However, within the subgroups there appear to be much greater levels of similarity, with 94% amino acid identity reported between two subgroup A strains, A2 and Long (Johnson et al., 1987), and 98% amino acid identity between two subgroup B strains, 18537 and 8/60 (Sullender et al., 1990).

We have previously reported the presence of a number of distinct lineages of RS virus during a single epidemic (Cane & Pringle, 1991). These lineages were defined by restriction mapping or sequencing parts of the nucleocapsid (N) protein and small hydrophobic (SH) protein genes. We now report the G protein gene nucleotide sequences of strains representing each of the subgroup A lineages. The deduced amino acid similarities of the G proteins of these isolates varied between 80% and 99%. In addition, we show that the amino acid variability was not evenly distributed over the G protein; in the extracellular part of the protein there was a sharply defined hypervariable domain separated from a more extended variable domain by a highly conserved region. These results show that a pattern of variability exists within subgroup A RS viruses similar to that observed by Johnson et al. (1987) in their comparison of the G proteins of the prototype strains of subgroups A and B.

Methods

Cells and viruses. RS virus strains were propagated in HEp-2 or MRC-5 cells. Virus strains examined included RSB642, RSB1734, RSB5857, RSB6190, RSB6256 and RSB6614, which were isolated in the Birmingham, U.K. area in 1989 as previously described (Cane & Pringle, 1991).

RNA extraction. RNA was extracted from infected cells by the method of Kumar & Lindberg (1972).
Polymerase chain reaction (PCR). cDNA synthesis was carried out as previously described (Cane & Pringle, 1991) using primer 1 (described below). A sample (5 μl) of cDNA was then amplified using primers 1 and 2 in a 100 μl reaction using VENT polymerase (New England Biolabs) according to the manufacturer's protocol.

PCR primers. Primer 1, 5′ (GGCCCGGGAAGC)TTTTTTTTTTTTTTTTT 3′, is a universal primer for polyadenylated mRNA (P. Chambers, personal communication). Primer 2, 5′ (GGATCCC)-GGGCCGAATTCGACATGTCG 3′, is homologous to nucleotides 1 to 21 of the G gene sequence of strain A2 (Wertz et al., 1985). The brackets indicate linkers.

Cloning and sequencing. PCR products were extracted as previously described (Cane & Pringle, 1991). Cloning was facilitated by the presence of a PstI site in the PCR products of G protein genes of most subgroup A RS viruses. G protein gene products were cloned in two parts, following Aal and PstI restriction digestion, into similarly digested Bluescribe vector (Vector Cloning Systems). Plasmid DNA was extracted from small-scale cultures by alkaline lysis for sequencing. The DNA was ethanol-precipitated following RNase treatment and phenol–chloroform extraction. The DNA was denatured with alkali, purified by centrifugation through a Sepharose CL4B column and immediately annealed with sequencing primer (forward or reverse). Sequencing reactions were carried out using a Sequenase kit (U.S. Biochemicals). An additional primer, 5′ GCAATCATACAAGATGC GGGGCAAATGCAAACATGTCC 3′, is homologous to nucleotides 1 to 21 of the G gene sequence of strain A2 (Wertz et al., 1985). The brackets indicate linkers.

Cloning and sequencing. PCR products were extracted as previously described (Cane & Pringle, 1991). Cloning was facilitated by the presence of a PstI site in the PCR products of G protein genes of most subgroup A RS viruses. G protein gene products were cloned in two parts, following Aal and PstI restriction digestion, into similarly digested Bluescribe vector (Vector Cloning Systems). Plasmid DNA was extracted from small-scale cultures by alkaline lysis for sequencing. The DNA was ethanol-precipitated following RNase treatment and phenol–chloroform extraction. The DNA was denatured with alkali, purified by centrifugation through a Sepharose CL4B column and immediately annealed with sequencing primer (forward or reverse). Sequencing reactions were carried out using a Sequenase kit (U.S. Biochemicals). An additional primer, 5′ GCAATCATACAAGATGC GGGGCAAATGCAAACATGTCC 3′, is homologous to nucleotides 1 to 21 of the G gene sequence of strain A2 (Wertz et al., 1985). The brackets indicate linkers.

Cloning and sequencing. PCR products were extracted as previously described (Cane & Pringle, 1991). Cloning was facilitated by the presence of a PstI site in the PCR products of G protein genes of most subgroup A RS viruses. G protein gene products were cloned in two parts, following Aal and PstI restriction digestion, into similarly digested Bluescribe vector (Vector Cloning Systems). Plasmid DNA was extracted from small-scale cultures by alkaline lysis for sequencing. The DNA was ethanol-precipitated following RNase treatment and phenol–chloroform extraction. The DNA was denatured with alkali, purified by centrifugation through a Sepharose CL4B column and immediately annealed with sequencing primer (forward or reverse). Sequencing reactions were carried out using a Sequenase kit (U.S. Biochemicals). An additional primer, 5′ GCAATCATACAAGATGC GGGGCAAATGCAAACATGTCC 3′, is homologous to nucleotides 1 to 21 of the G gene sequence of strain A2 (Wertz et al., 1985). The brackets indicate linkers.

Results

The nucleotide sequences of the mRNAs of the G protein genes of strains RSB5857 representing lineage SHL1, RSB6190 representing lineage SHL2, RSB6256 representing lineage SHL3, RSB6614 representing lineage SHL4, and RSB642 and RSB1734 representing lineage SHL5 are shown in Fig. 1, together with those of the prototype subgroup A strains Long and A2 (Wertz et al., 1985; Johnson et al., 1987). All strains examined had 917 nucleotides, as observed for Long, one nucleotide less than for A2. The nucleotide homology between the strains is shown in Table 1. In every case the degree of amino acid similarity was less than the nucleotide similarity. Strains RSB642, RSB1734, RSB5857, RSB6256 and RSB6614 all lacked the carboxy-terminal amino acid observed in the SHL2 lineage strain RSB6190 and the prototype strains A2 and Long, which were previously considered to be most closely related to lineage SHL2. Amino acid similarities ranged between 80% and 99%, the least similar strains being RSB6190 and RSB642, and the most similar being RSB5857 and RSB6256. Fig. 2(b) is a dendrogram showing the relatedness of the various strains as determined by the amino acid sequences of the G proteins. It shows a distribution very similar to that given by the dendrogram based on nucleotide sequences, except that Long and A2 are slightly further apart.

The relationships between the strains established in this way correspond well with the relationships between the five lineages previously defined by analysis of the SH and N protein genes (Cane & Pringle, 1991). The only exception is that RSB6190 (SHL2) appears, by analysis of the G protein gene, to be more closely related to RSB5857, RSB6256 and RSB6614, whereas in the SH and N protein gene analyses it appeared closer to Long and A2, with which it shares the extra terminal amino acid.

A striking feature of the amino acid sequences of the G proteins is the large number of threonine residues which occur as pairs. For example, in RSB6614 there are 22 single threonines but 39 threonines in groups of two or more. In addition, in the region of the protein between amino acid residues 193 and 240, there are five repetitions of the sequence ..KPXnTTKXn... The significance of these features is not known, but may be associated with the extensive O-glycosylation of the protein, or some essential structural feature.

The amino acid variability was not evenly distributed throughout the protein. Fig. 4 is a plot of the extent of heterology of the G proteins of all the strains shown in Fig. 3; groups of 10 amino acids each overlapping by five amino acids have been analysed for each point. From
RS virus subgroup A G protein variability

Fig. 1. Nucleotide sequences of mRNAs of G protein genes. The sequences for strains Long and A2 have been taken from Johnson et al. (1987). Nucleotides 1 to 21 were determined by using PCR primer 1. The asterisks indicate the positions of mismatches. The positions of the initiating ATG and terminating TAG codons are underlined.

this plot and from Fig. 3, it can be seen that the 67 amino acids at the amino terminus are conserved, followed by amino acids 67 to 74 which are fairly variable. There is then another small conserved area, followed by a highly variable domain between amino acids 101 and 133. From amino acid 147 to amino acid 207, there is almost total conservation, followed by a fairly variable carboxy terminus.

Clearer examination of the highly variable domain bounded by amino acids 101 and 133 shows that this area is variable between all the strains examined. The number of amino acid differences in this region varied between two (6%) and 16 (48%). Most nucleotide differences in the part of the G protein gene coding for this area (nucleotides 316 to 414) are first or second position changes in the codon triplets. This highly variable domain is bounded by two conserved N-glycosylation sites at amino acid residues 85 and 135. Within this highly variable domain two amino acid triplets, TTP at position 118 to 120, and TTG at position 129 to 131 are conserved. This domain corresponds to a hydrophilic peak observed in the hydrophilicity plot of the G protein (not shown).
all the strains into consideration, there are a further six possible N-glycosylation sites, of which between one and five are available depending on the strain. The positions of these sites are indicated in Fig. 3.

Three N-glycosylation sites are conserved over the entire G protein, i.e. those mentioned above at amino acids 85 and 135, and another at amino acid 237. Taking all the strains into consideration, there are a further six possible N-glycosylation sites, of which between one and five are available depending on the strain. The positions of these sites are indicated in Fig. 3.

Discussion

This paper reports the nucleotide and deduced amino acid sequences of the G protein genes of six subgroup A RS virus strains isolated in 1989 and compares these data with the sequences from two prototype strains. Analysis of the relatedness of the nucleotide sequences of the G protein genes generated a dendrogram (Fig. 2a) similar to that generated by CLUSTAL (Higgins & Sharp, 1988) to show the relatedness of nucleotide sequences of G protein genes. (b) Dendrogram to show the relatedness of G protein amino acid sequences.

Fig. 3. Deduced amino acid sequences of G proteins. The sequences of the Long and A2 strains have been taken from Johnson et al. (1987). Symbols: * indicates non-conservative changes; + indicates conservative changes; • indicates conserved N-glycosylation sites; <> indicates non-conserved N-glycosylation sites; • indicates possible cleavage sites for the production of soluble G protein (Hendricks et al., 1988). The presumed hydrophobic anchor is overlined.

Fig. 4. Histogram to show distribution of variable amino acids occurring in varying numbers of strains.
to that based on about 150 nucleotides at the 5' end of the SH protein genes determined previously (Cane & Pringle, 1991). This result confirms the validity of the procedure of dividing the isolates into distinct lineages based on a much shorter run of mainly non-coding sequence, which is unlikely to be subject to immunological selective pressure. The broader divisions of the 1989 isolates (Fig. 2a), namely RSB642 with RSB1734, RSB5857 with RSB6256 and RSB6614, and RSB6190, reflected the different N protein gene restriction patterns (NP2, NP4 and NP5) previously determined (Cane & Pringle, 1991).

The extent of amino acid heterogeneity between the 1989 subgroup A strains is considerably greater (up to 20%) than that previously reported for the two prototype subgroup A strains, A2 and Long (6%) (Johnson et al., 1987), or between two subgroup B strains (2%) (Sullender et al., 1990). The results with the subgroup A strains correlate with our previous suggestion that the two prototype laboratory strains, Long and A2, belong to a lineage which may be related to SIHL2, but which represent only a biased sample of the subgroup A strains in circulation (Cane & Pringle, 1991). The greater extent of heterogeneity reported here is comparable with that seen when comparing the HA2 region of the haemagglutinin protein of HI and H2 subtypes of influenza A virus (reviewed by Smith & Palese, 1989).

The variation in inferred amino acid sequence is not spread evenly throughout the RS virus G protein. The conserved presumptive cytoplasmic and transmembrane regions are followed by a sharply defined hypervariable domain between amino acid residues 101 and 133. A more extended variable region is located in the carboxy-terminal third of the protein, which may be made up of three smaller variable domains (Fig. 4). These two major variable regions are separated by a highly conserved domain between amino acids 147 and 207. Since most of the nucleotide changes in the parts of the gene coding for the variable areas result in coding changes, as previously described in a comparison of G protein genes from prototype strains of subgroups A and B (Johnson & Collins, 1989), there may be immunological pressure for change in these regions. The division of the G protein into a conserved cytoplasmic and transmembrane region, together with a conserved central region, is also very similar to the pattern given by comparison of the G proteins from prototype strains from subgroups A and B (Johnson et al., 1987). However, the central conserved region of the subgroup A strains described here is much larger and extends for 51 amino acids (from residues 147 to 207) as compared to the 13 amino acids (from residues 164 to 176) conserved between the subgroups.

Localization of amino acid variability is a feature of the influenza A virus haemagglutinin (Both et al., 1983), and the variable regions of the protein correlate with antigenic drift sites (Wiley et al., 1981). It is possible that this may also be the case for RS virus G protein. If this were so, one would predict antigenic drift sites at approximate amino acid positions 67 to 74, 100 to 135, 220 to 235, 240 to 260 and 280 to the end of the gene. Norby et al. (1987) used synthetic peptides for analysis of the antigenicity of the G protein by site-specific serology and found that rabbit hyperimmune antisera raised against some subgroup A strains reacted with peptides representing amino acid residues 84 to 108, 134 to 158, 174 to 188, 194 to 208 and 224 to 238 of strain A2, whereas mouse MAbs reacted only with a peptide representing amino acids 174 to 188, and human convalescent sera with peptides representing amino acids 144 to 158 and 174 to 208. It seems that the human sera and mouse MAbs recognized only the highly conserved areas of the protein (namely residues 134 to 158, 174 to 188 and 194 to 208). The failure of the human sera to recognize the other residues recognized by the rabbits, namely 84 to 108 and 224 to 238, may have been due to a difference in the specificity of the response as a consequence of infection by a heterologous virus. However, the conclusions that can be drawn from these peptide studies are limited since many epitopes on the G protein may be conformational rather than linear.

The importance of the carboxy-terminal third of the G protein in the immune response has been examined by others. Using vaccinia virus recombinants expressing progressively truncated portions of the G protein, it was demonstrated that a recombinant G230 protein which lacked the terminal 68 amino acids, but not G180 which lacked the terminal 118 amino acids, was as effective at inducing a neutralizing antibody response as recombinants expressing the complete G protein (Olmsted et al., 1989). However, Garcia-Barreno et al. (1990), using neutralization escape mutants, found that as a result of a frameshift mutation that changed all the amino acids after residue 205, the G protein of the mutant strains failed to react with most anti-G MAbs or with polyclonal antisera, indicating that the carboxy-terminal third is essential for the integrity of many epitopes. Taking these findings together, it appears that the region of the G protein between amino acids 205 and 230, shown to be partly variable in this study, may include important determinants for the induction of neutralizing antibodies.

It has been shown that both N- and O-glycosylation of the G protein affects antibody binding to the protein (Palomo et al., 1991), and similar effects have been demonstrated previously for influenza A virus (Skehel et al., 1984). The difference in the number of N-glycosylation sites in the strains described in this report, particularly in the variable (and presumably immunolo-
The significance of the variability of the G protein in the epidemiology of RS virus remains to be established. In this report we have demonstrated that certain areas of subgroup A is at least as great as that seen during prolonged antigenic drift of influenza A virus (Both et al., 1983). The subgroup A RS virus strains isolated during a single year in a well defined geographical area could be divided into three main lineages, with some reduced variation within these lineages. Whether the variation within the lineages represents antigenic drift or non-progressive random variation are questions that can only be answered by analysis of successive annual epidemics in the same population.

The authors wish to thank Dr U. Desselberger and Mrs R. Tudor of the Regional Virus Laboratory, East Birmingham Hospital for collecting the RS virus strains used in this study. Thanks also to Drs A. Easton and M. McCrae for help with the computer analyses. This work was supported by MRC programme grant no. PG8322715.

References


JOHNSON, P. R. & COLLINS, P. L. (1989). The 1B (NS2), 1C (NS1) and N proteins of human respiratory syncytial virus (RSV) of antigenic subgroups A and B: sequence conservation and divergence within RSV genomic RNA. Journal of General Virology 70, 1539–1547.


MOSS, B. & COLE, P. A. Cane, D. A. Matthews and C. R. Pringle

National Virology 151, 626–633.


MOSS, B. & COLE, P. A. Cane, D. A. Matthews and C. R. Pringle

National Virology 151, 626–633.


MOSS, B. & COLE, P. A. Cane, D. A. Matthews and C. R. Pringle

National Virology 151, 626–633.


MOSS, B. & COLE, P. A. Cane, D. A. Matthews and C. R. Pringle

National Virology 151, 626–633.