scientific reports

OPEN

Respiratory syncytial virus B sequence analysis reveals a novel early genotype

Juan C. Muñoz-Escalante[1], Andreu Comas-García[1,2], Sofía Bernal-Silva[1,2] & Daniel E. Noyola[1]✉

Respiratory syncytial virus (RSV) is a major cause of respiratory infections and is classified in two main groups, RSV-A and RSV-B, with multiple genotypes within each of them. For RSV-B, more than 30 genotypes have been described, without consensus on their definition. The lack of genotype assignation criteria has a direct impact on viral evolution understanding, development of viral detection methods as well as vaccines design. Here we analyzed the totality of complete RSV-B G gene ectodomain sequences published in GenBank until September 2018 (n = 2190) including 478 complete genome sequences using maximum likelihood and Bayesian phylogenetic analyses, as well as intergenotypic and intragenotypic distance matrices, in order to generate a systematic genotype assignation. Individual RSV-B genes were also assessed using maximum likelihood phylogenetic analyses and multiple sequence alignments were used to identify molecular markers associated to specific genotypes. Analyses of the complete G gene ectodomain region, sequences clustering patterns, and the presence of molecular markers of each individual gene indicate that the 37 previously described genotypes can be classified into fifteen distinct genotypes: BA, BA-C, BA-CC, CB1-THB, GB1-GB4, GB6, JAB1-NZB2, SAB1, SAB2, SAB4, URU2 and a novel early circulating genotype characterized in the present study and designated GB0.

Respiratory syncytial virus (RSV) is a leading cause of lower respiratory tract infections in infants, elderly adults, and immunosuppressed individuals[1]. Since the discovery of RSV, a wide diversity of viral strains has been identified leading to the classification in two major groups (RSV-A and RSV-B), as well as multiple genotypes[2–4]. RSV infections occur worldwide and co-circulation of viral strains from both major groups is common[5]. RSV-B strains are the predominant viruses in approximately one third of winter seasons[6,7]. Since the initial description of RSV genotypes, there has been an increasing number of reported genotypes, with worldwide extension of novel strains and apparent extinction of older types. The diversity in genotypes of RSV-B strains is greater than for RSV A. While RSV-A strains can be grouped into seven distinct genotypes, there have been at least 37 RSV-B genotypes described in the literature (GB1, GB2, GB3, GB4, GB5, GB6, GB12, GB13, SAB1, SAB2, SAB3, SAB4, URU1, URU2, CB1, THB, BA1, BA2, BA3, BA4, BA5, BA6, BA7, BA8, BA9, BA10, BA11, BA12, BA13, BA14, BA-Ly, BA-C, BA-CCA, BA-CCB, JAB1, NZB1, and NZB2)[4,9–27]. Of note, at present there is no consensus regarding criteria to discriminate between genotypes[4,8,17,20,28,29]. Identification and description of many genotypes has relied on sequencing of the second hypervariable region of the G gene; however, analyses limited to this region are not always able to distinguish between strains that might be considered as part of the same or different genotype[8,30]. As a result, some viral clusters which have been described as distinct genotypes have turned out to belong to previously identified genotypes. Classification of viral strains is of relevance not only for taxonomic purposes, but in order to better understand the epidemiology of this important virus, as well as the development of therapeutic and preventive strategies. In the present work we have analyzed a large set of RSV-B sequences using a methodology previously described for RSV-A genotype analysis[8]. Our results indicate that many of the 37 previously described genotypes can be reorganized within a smaller number of genotypes based on intra and inter-clade variability. In addition, a previously unrecognized RSV-B cluster composed of strains that circulated between 1972 and 1983 shows unique characteristics enough to identify them as a distinct early genotype.

[1]Microbiology Department, Facultad de Medicina, Universidad Autónoma de San Luis Potosí, Avenida Venustiano Carranza 2405, San Luis Potosí 78210, México. [2]Center for Research in Biomedicine and Health Sciences, Facultad de Medicina, Universidad Autónoma de San Luis Potosí, San Luis Potosí, Mexico. ✉email: dnoyola@uaslp.mx

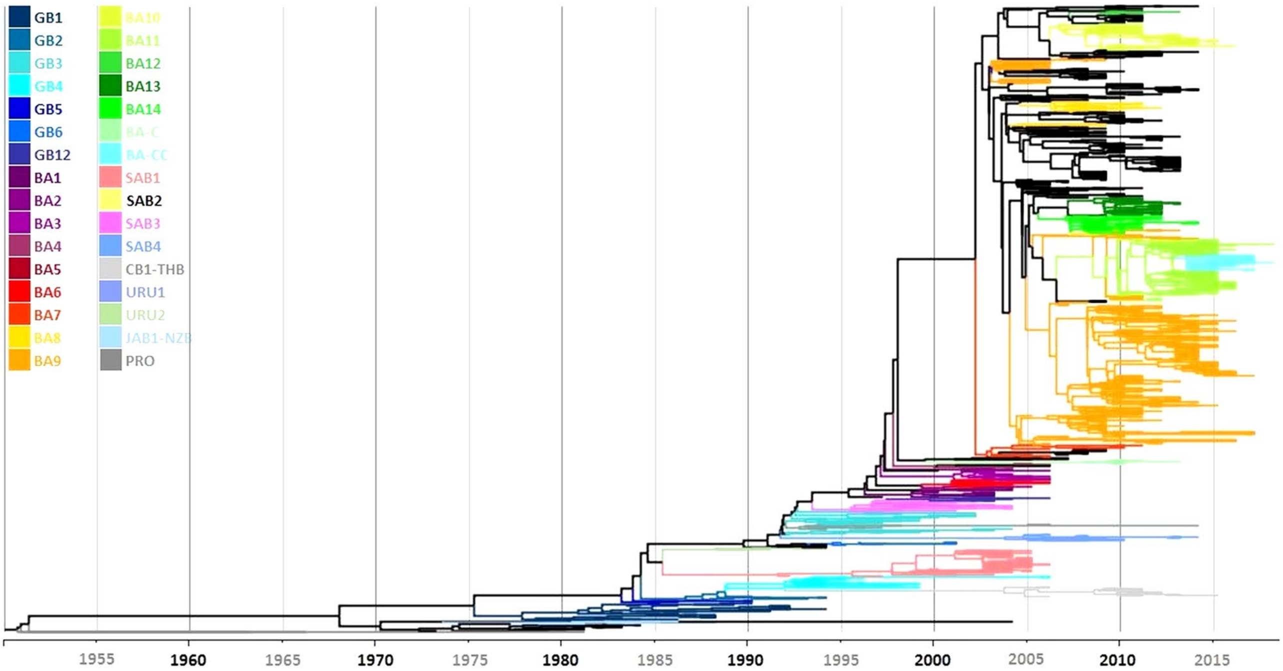


Figure 1. Phylogenetic tree of 1,334 unique RSV-B complete G gene ectodomain sequences constructed by Bayesian MCMC analysis. Genotype assignment was carried out with the use of 169 reference sequences including 37 previously described genotypes and prototype strains.

Results

Dataset selection.

From a total of 10,340 RSV sequences downloaded from NCBI, 3029 corresponded to the RSV-B complete G gene ectodomain; nevertheless, 831 sequences (27.4%) were discarded due to indels or degenerate nucleotides that could interfere in a correct genotype assignation (Supplementary Fig. 1). The final dataset comprised 2198 sequences with at least the complete G gene ectodomain and was used for genotypes analysis. This dataset included 1673 sequences of only complete G gene ectodomain, 18 sequences (0.82%) of only the complete G gene, 29 sequences (1.32%) of only SH and G genes, and 478 complete genome sequences. For all the sequences, the complete G gene ectodomain was used to assign genotypes; sequences with one or more complete genes were used for cladistic analyses and detection of molecular markers at nucleotide and amino acid sequences.

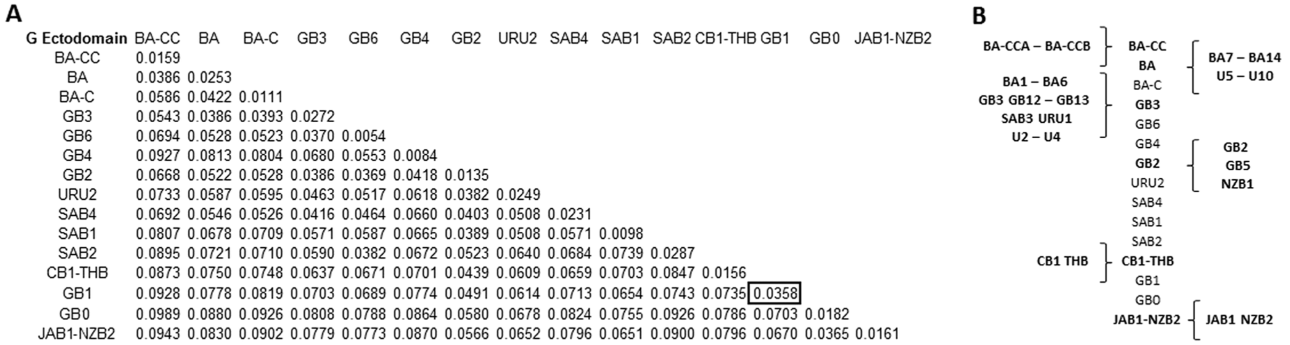
Genotype assignment.

Genotype assignment was carried out by clustering of 1,334 unique RSV-B complete G gene ectodomain sequences (corresponding to 60.7% of the dataset) with sequences previously designated as reference sequences or with equivalent reference sequences, defined as described in the Methods section, using both Maximum Likelihood method and Bayesian MCMC. Clade distribution, topology and clustering of sequences were concordant in both methods (Fig. 1 and Supplementary Fig. 2).

The largest number of sequences corresponded to the BA9 genotype (27.61%), followed by SAB1 (7.11%), and BA11 (5.05%); all other genotypes contributed with less than 2.5% each. Of note, CB1 and THB genotypes previously described by Cui et al.[15] and Auksornkitti et al.[25], JAB1 and NZB2 previously described by Kuroiwa et al.[18] and Matheson et al.[19], and BA-CCA and BA-CCB genotypes described by Gaymard et al.[13] clustered and intermingled in individual clades, suggesting that each couple of genotypes correspond to the same genotype (hereafter referred to as CB1-THB, JAB1-NZB2, and BA-CC), which was corroborated during molecular marker analysis as described below. Ten well defined and sustained clades including two or more sequences did not cluster with any reference or equivalent reference sequences and were assigned as unidentified clades (U1-10); all these clades, except U1, had the 60-nucleotide duplication characteristic of BA strains. Eight sequences did not cluster with any other sequence either within previously described genotypes or unidentified clades; because of this, they were considered as singletons and were excluded from subsequent genotype analyses.

To corroborate the genotype assignment, an intergenotypic and intragenotypic p-distance matrix was generated with all the sequences (n = 2190) which were assigned to previously described genotypes (n = 37) and unidentified clades (n = 10). GB1 presented the highest intragenotypic distance (p = 0.0358) and this value was used as the threshold to identify clades which belong to the same or different genotype (Supplementary Fig. 3).

Genotypes and unidentified clades were grouped following a stepwise lowest distance neighbor joining strategy until all groups distance were higher than the threshold. This resulted in the joining of genotypes GB2, GB5 and NZB1 in a single genotype designated as genotype GB2; the joining of genotypes GB3, GB12, GB13, SAB3, URU1, BA1-6 and clades U2-4 into a single genotype designated as genotype GB3; and the joining of genotypes BA7-14 and clades U5-U10, designated as genotype BA (Fig. 2). Remarkably, an independent clade (U1) with sequences from strains isolated up to seven years (1972) before isolation of the first GB1 sequence included in


A
B
BA-CCA – BA-CCB
BA1 – BA6
GB3 GB12 – GB13
SAB3 URU1
U2 – U4
CB1 THB
BA-CC
BA
BA-C
GB3
GB6
GB4
GB2
URU2
SAB4
SAB1
SAB2
CB1-THB
GB1
GB0
JAB1-NZB2
BA7 – BA14
U5 – U10
GB2
GB5
NZB1
JAB1 NZB2

G Ectodomain	BA-CC	BA	BA-C	GB3	GB6	GB4	GB2	URU2	SAB4	SAB1	SAB2	CB1-THB	GB1	GB0	JAB1-NZB2
BA-CC	0.0159														
BA	0.0386	0.0253													
BA-C	0.0586	0.0422	0.0111												
GB3	0.0543	0.0386	0.0393	0.0272											
GB6	0.0694	0.0528	0.0523	0.0370	0.0054										
GB4	0.0927	0.0813	0.0804	0.0680	0.0553	0.0084									
GB2	0.0668	0.0522	0.0528	0.0386	0.0369	0.0418	0.0135								
URU2	0.0733	0.0587	0.0595	0.0463	0.0517	0.0618	0.0382	0.0249							
SAB4	0.0692	0.0546	0.0526	0.0416	0.0464	0.0660	0.0403	0.0508	0.0231						
SAB1	0.0807	0.0678	0.0709	0.0571	0.0587	0.0665	0.0389	0.0508	0.0571	0.0098					
SAB2	0.0895	0.0721	0.0710	0.0590	0.0382	0.0672	0.0523	0.0640	0.0684	0.0739	0.0287				
CB1-THB	0.0873	0.0750	0.0748	0.0637	0.0671	0.0701	0.0439	0.0609	0.0659	0.0703	0.0847	0.0156			
GB1	0.0928	0.0778	0.0819	0.0703	0.0689	0.0774	0.0491	0.0614	0.0713	0.0654	0.0743	0.0735	$\boxed{0.0358}$		
GB0	0.0989	0.0880	0.0926	0.0808	0.0788	0.0864	0.0580	0.0678	0.0824	0.0755	0.0926	0.0786	0.0703	0.0182	
JAB1-NZB2	0.0943	0.0830	0.0902	0.0779	0.0773	0.0870	0.0566	0.0652	0.0796	0.0651	0.0900	0.0796	0.0670	0.0365	0.0161

Figure 2. (**A**) Distinct genotypes identified through intergenotypic and intragenotypic p-distance analysis of 2,190 RSVB complete ectodomain sequences. The highest intragenotypic distance was observed for GB1 (0.0358). All clusters with intergenotypic distance higher than this threshold value were considered as distinct genotypes. (**B**) Several previously described genotypes or unique unidentified clusters were found to cluster together with BA-CC (BA-CCA and BA-CCB), CB1-THB (CB1 and THB), JAB1-NZB2 (JAB1 and NZB2), GB2 (GB2, GB5, and NZB1), GB3 (GB3, GB12-GB13, SAB3, URU1, BA1-BA6, U2-U4), and BA (BA7-BA14, U5-U10).

the dataset (1979) was identified; this clade was subsequently assigned as genotype GB0 (Fig. 3; Supplementary Fig. 4). The lowest intergenotypic p-distance between this clade and the rest of genotypes was 0.0578 (when compared with GB2), a value exceeding by 1.6 times the threshold value of 0.0358 (Fig. 2). Sequences included in the GB0 genotype clade have not been described to conform a unique genotype previously.

Complete NS1, NS2, N, P, M, SH, G, F, M2, and L gene analysis.

For each of the ten RSV-B genes (NS1, NS2, N, P, M, SH, G, F, M2, and L) cladograms were generated from the corresponding Maximum Likelihood analysis under the best fitting substitution model for each gene dataset, as well as the corresponding intergenotypic and intragenotypic p-distance matrices based on the previously assigned genotypes (Fig. 4). Genotypes SAB2 and CB1-THB were not included on individual gene analysis due to lack of complete genes or complete genomes sequences for these genotypes; URU2 was only included on complete SH and G genes analysis due the presence of only partial genome sequences for this genotype.

Cladogram topologies and sequence clustering was concordant in the majority of genes; sequences assigned to a specific genotype grouped in well differentiated clusters, with exception of the recently identified genotype designated as BA-CC. Sequences assigned as BA-CC grouped on two different but proximate clades in the NS1 gene cladogram. Furthermore, p-distance value analysis showed concordant genotype assignation for most genotypes in most gene matrices, with NS1 matrix being the exception with 19 discordances. Overall, 50 (7.3%) of the 684 intergenotypic comparisons had p values lower than the threshold. This was partly explained by the small number of sequences available for some genotypes; for instance, for comparisons for which there were 20 sequences or less the proportion of intergenotypic p-values lower than the threshold was higher (30 of 275 instances, 10.9%) than for comparisons for which there were more than 20 sequences available (20 of 409 instances, 4.9%; $P = 0.003$). Of note, GB6 (for which there was only one full genome sequence) was included as one of the genotypes in 20 (40%) of the 50 comparisons in which the p-value was below the threshold.

All RSV genes datasets were assessed for recombination with RDP, GENECONV, Chimaera, MAxChi, BootScan, SiScan and 3Seq algorithms using RDP4 v.4.100[31], as well as GARD algorithm[32]. There was no evidence of recombination among RSV sequences included in the study.

Molecular markers detection.

For each of the ten RSV-B genes (NS1, NS2, N, P, M, SH, G, F, M2, and L), nucleotide sequences spanning from 3′UTR to 5′UTR were aligned and grouped in accordance with genotype assignment. Each genotype was compared against RSV-B reference sequence "strain B1" (Accession Number NC_001781.1) and every variant at every site was recorded. Amino acid sequences were deduced from each of the ten RSV-B coding regions, and variants were recorded as previously described[8]. Variants fixed in more than 75% of the genotype sequences were considered as molecular markers.

In total, 1,213 nucleotide variants distributed at the total length of the genome fulfilled the criteria to be considered molecular markers; 636 (52.4%) of them were present in a single genotype (Fig. 5 and Table 1). In addition, 213 deduced amino acid variants at the total of proteins fulfilled the criteria of molecular markers; 107 (50.2%) of these molecular markers were present in a single genotype. For genotypes GB2, GB3, and BA no unique molecular markers were detected. Genotype GB0 had 71 nucleotide molecular markers and 8 amino acid molecular markers which were unique for this genotype (Fig. 5, Table 1, and Supplementary Table 1).

Geographic and temporal distribution.

Date and country (continent) of isolation of the strain corresponding to each sequence in the dataset was recorded. As noted previously, eight (0.36%) of the 2198 sequences in the dataset were not assigned to any genotype. Geo-temporal records showed circulation of GB0 starting in 1972, seven years later than the last Prototype RSV-B sequence was isolated in Europe. This was the only genotype detected up to 1979, when GB1 was first isolated; these two genotypes co-circulated until the mid-80s.

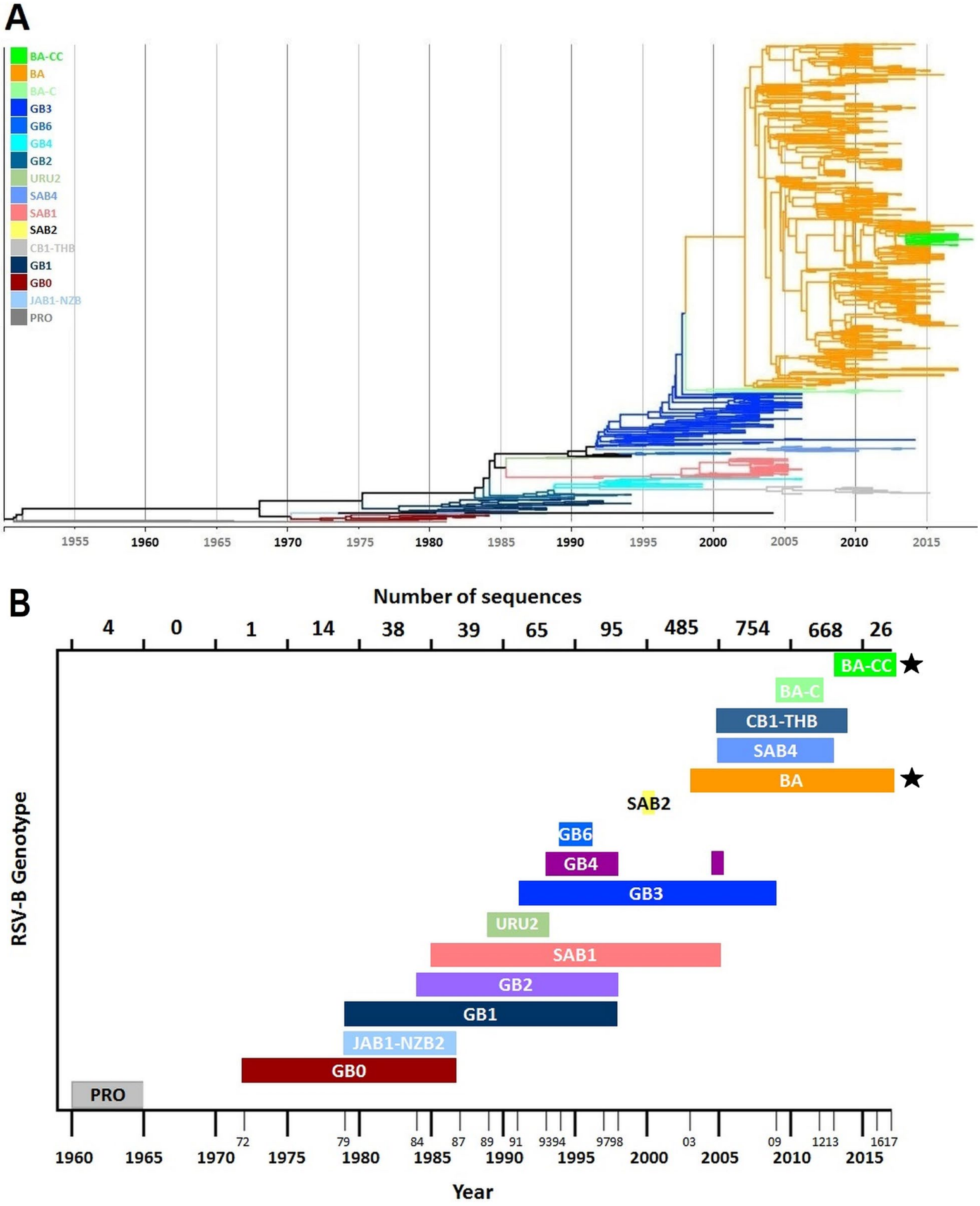
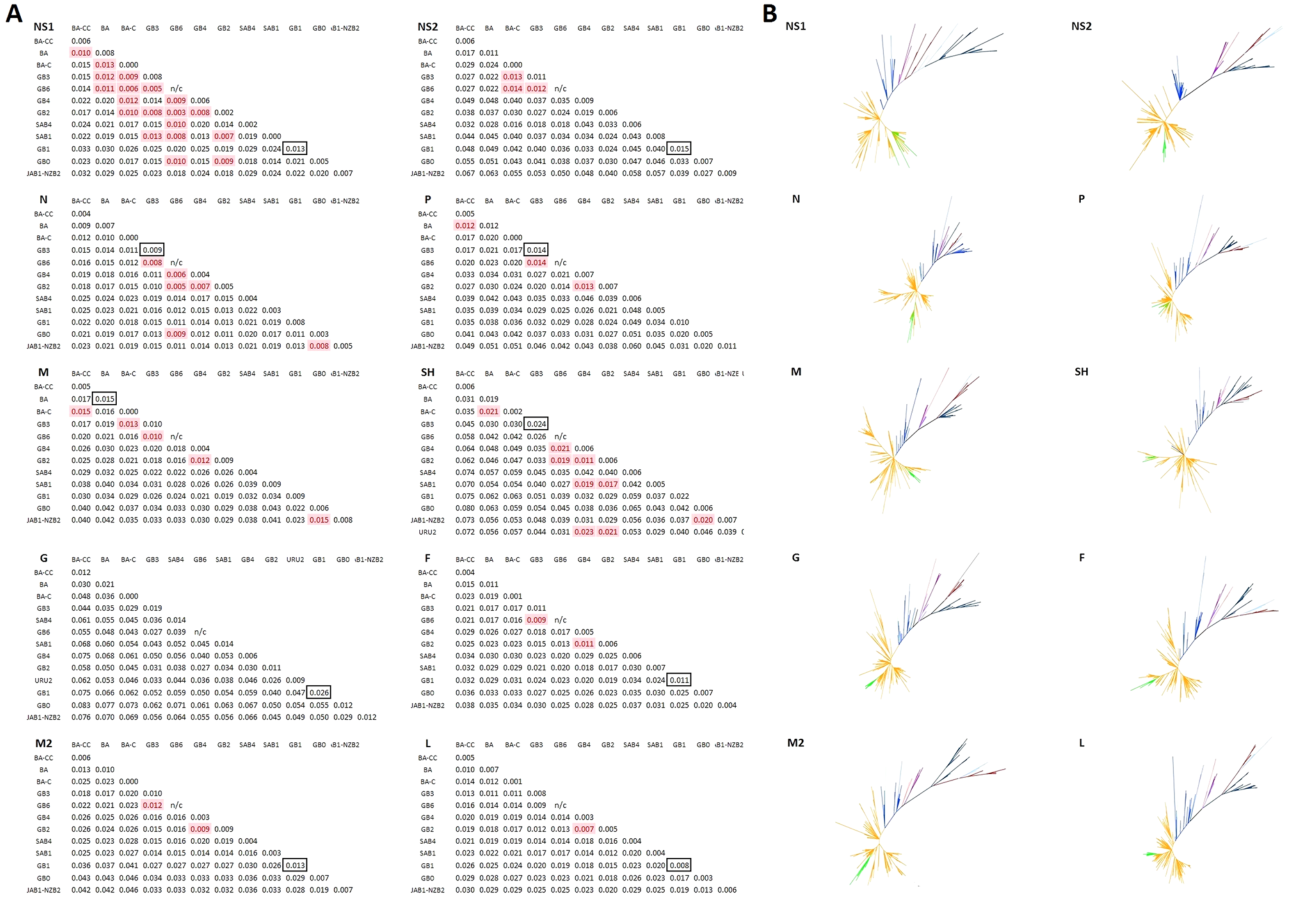


Figure 3. (**A**) Phylogenetic tree of 1,334 unique RSV-B complete G gene ectodomain sequences constructed by Bayesian MCMC analysis showing the 15 distinct genotypes defined through intragenotypic and intergenotypic p-distance analysis. (**B**) Temporal distribution of RSV-B genotypes since their first up to their last detection. Genotypes marked with stars indicate genotypes currently in circulation.



NS1	BA-CC	BA	BA-C	GB3	GB6	GB4	GB2	SAB4	SAB1	GB1	GB0	B1-NZB2
BA-CC	0.006											
BA	0.010	0.008										
BA-C	0.015	0.013	0.000									
GB3	0.015	0.012	0.009	0.008								
GB6	0.014	0.011	0.006	0.005	n/c							
GB4	0.022	0.020	0.012	0.014	0.009	0.006						
GB2	0.017	0.014	0.010	0.008	0.003	0.008	0.002					
SAB4	0.024	0.021	0.017	0.015	0.010	0.020	0.014	0.002				
SAB1	0.022	0.019	0.015	0.013	0.008	0.013	0.007	0.019	0.000			
GB1	0.033	0.030	0.026	0.025	0.020	0.025	0.019	0.029	0.024	0.013		
GB0	0.023	0.020	0.017	0.015	0.010	0.015	0.009	0.018	0.014	0.021	0.005	
JAB1-NZB2	0.032	0.029	0.025	0.023	0.018	0.024	0.018	0.029	0.024	0.022	0.020	0.007

NS2	BA-CC	BA	BA-C	GB3	GB6	GB4	GB2	SAB4	SAB1	GB1	GB0	B1-NZB2
BA-CC	0.006											
BA	0.017	0.011										
BA-C	0.029	0.024	0.000									
GB3	0.027	0.022	0.013	0.011								
GB6	0.027	0.022	0.014	0.012	n/c							
GB4	0.049	0.048	0.040	0.037	0.035	0.009						
GB2	0.038	0.037	0.030	0.027	0.024	0.019	0.006					
SAB4	0.032	0.028	0.016	0.018	0.018	0.043	0.033	0.006				
SAB1	0.044	0.045	0.040	0.037	0.034	0.034	0.024	0.043	0.008			
GB1	0.048	0.049	0.042	0.040	0.036	0.033	0.024	0.045	0.040	0.015		
GB0	0.055	0.051	0.043	0.041	0.038	0.037	0.030	0.047	0.046	0.033	0.007	
JAB1-NZB2	0.067	0.063	0.055	0.053	0.050	0.048	0.040	0.058	0.057	0.039	0.027	0.009

N	BA-CC	BA	BA-C	GB3	GB6	GB4	GB2	SAB4	SAB1	GB1	GB0	B1-NZB2
BA-CC	0.004											
BA	0.009	0.007										
BA-C	0.012	0.010	0.000									
GB3	0.015	0.014	0.011	0.009								
GB6	0.016	0.015	0.012	0.008	n/c							
GB4	0.019	0.018	0.016	0.011	0.006	0.004						
GB2	0.018	0.017	0.015	0.010	0.005	0.007	0.005					
SAB4	0.025	0.024	0.023	0.019	0.014	0.017	0.015	0.004				
SAB1	0.025	0.023	0.021	0.016	0.012	0.015	0.013	0.022	0.003			
GB1	0.022	0.020	0.018	0.015	0.011	0.014	0.013	0.021	0.019	0.008		
GB0	0.021	0.019	0.017	0.013	0.009	0.012	0.011	0.020	0.017	0.011	0.003	
JAB1-NZB2	0.023	0.021	0.019	0.015	0.011	0.014	0.013	0.021	0.019	0.013	0.008	0.005

P	BA-CC	BA	BA-C	GB3	GB6	GB4	GB2	SAB4	SAB1	GB1	GB0	B1-NZB2
BA-CC	0.005											
BA	0.012	0.012										
BA-C	0.017	0.020	0.000									
GB3	0.017	0.021	0.017	0.014								
GB6	0.020	0.023	0.020	0.014	n/c							
GB4	0.033	0.034	0.031	0.027	0.021	0.007						
GB2	0.027	0.030	0.024	0.020	0.014	0.013	0.007					
SAB4	0.039	0.042	0.043	0.035	0.033	0.046	0.039	0.006				
SAB1	0.035	0.039	0.034	0.029	0.025	0.026	0.021	0.048	0.005			
GB1	0.035	0.038	0.036	0.032	0.029	0.028	0.024	0.049	0.034	0.010		
GB0	0.041	0.043	0.042	0.037	0.033	0.031	0.027	0.051	0.035	0.020	0.005	
JAB1-NZB2	0.049	0.051	0.051	0.046	0.042	0.043	0.038	0.060	0.045	0.031	0.020	0.011

M	BA-CC	BA	BA-C	GB3	GB6	GB4	GB2	SAB4	SAB1	GB1	GB0	B1-NZB2
BA-CC	0.005											
BA	0.017	0.015										
BA-C	0.015	0.016	0.000									
GB3	0.017	0.019	0.013	0.010								
GB6	0.020	0.021	0.016	0.010	n/c							
GB4	0.026	0.030	0.023	0.020	0.018	0.004						
GB2	0.025	0.028	0.021	0.018	0.016	0.012	0.009					
SAB4	0.029	0.032	0.025	0.022	0.022	0.026	0.026	0.004				
SAB1	0.038	0.040	0.034	0.031	0.028	0.026	0.026	0.039	0.009			
GB1	0.030	0.034	0.029	0.026	0.024	0.021	0.019	0.032	0.034	0.009		
GB0	0.040	0.042	0.037	0.034	0.033	0.030	0.029	0.038	0.043	0.022	0.006	
JAB1-NZB2	0.040	0.042	0.035	0.033	0.033	0.030	0.029	0.038	0.041	0.023	0.015	0.008

SH	BA-CC	BA	BA-C	GB3	GB6	GB4	GB2	SAB4	SAB1	GB1	GB0	B1-NZE	(
BA-CC	0.006												
BA	0.031	0.019											
BA-C	0.035	0.021	0.002										
GB3	0.045	0.030	0.030	0.024									
GB6	0.058	0.042	0.042	0.026	n/c								
GB4	0.064	0.048	0.049	0.035	0.021	0.006							
GB2	0.062	0.046	0.047	0.033	0.019	0.011	0.006						
SAB4	0.074	0.057	0.059	0.045	0.035	0.042	0.040	0.006					
SAB1	0.070	0.054	0.054	0.040	0.027	0.019	0.017	0.042	0.005				
GB1	0.075	0.062	0.063	0.051	0.039	0.032	0.029	0.059	0.037	0.022			
GB0	0.080	0.063	0.059	0.054	0.045	0.038	0.036	0.065	0.043	0.042	0.006		
JAB1-NZB2	0.073	0.056	0.053	0.048	0.039	0.031	0.029	0.056	0.036	0.037	0.020	0.007	
URU2	0.072	0.056	0.057	0.044	0.031	0.023	0.021	0.053	0.029	0.040	0.046	0.039	(

G	BA-CC	BA	BA-C	GB3	SAB4	GB6	SAB1	GB4	GB2	URU2	GB1	GB0	B1-NZB2
BA-CC	0.012												
BA	0.030	0.021											
BA-C	0.048	0.036	0.000										
GB3	0.044	0.035	0.029	0.019									
SAB4	0.061	0.055	0.045	0.036	0.014								
GB6	0.055	0.048	0.043	0.027	0.039	n/c							
SAB1	0.068	0.060	0.054	0.043	0.052	0.045	0.014						
GB4	0.075	0.068	0.061	0.050	0.056	0.040	0.053	0.006					
GB2	0.058	0.050	0.045	0.031	0.038	0.027	0.034	0.030	0.011				
URU2	0.062	0.053	0.046	0.033	0.044	0.036	0.038	0.046	0.026	0.009			
GB1	0.075	0.066	0.062	0.052	0.059	0.050	0.054	0.059	0.040	0.047	0.026		
GB0	0.083	0.077	0.073	0.062	0.071	0.061	0.063	0.067	0.050	0.054	0.055	0.012	
JAB1-NZB2	0.076	0.070	0.069	0.056	0.064	0.055	0.056	0.066	0.045	0.049	0.050	0.029	0.012

F	BA-CC	BA	BA-C	GB3	GB6	GB4	GB2	SAB4	SAB1	GB1	GB0	B1-NZB2
BA-CC	0.004											
BA	0.015	0.011										
BA-C	0.023	0.019	0.001									
GB3	0.021	0.017	0.017	0.011								
GB6	0.021	0.017	0.016	0.009	n/c							
GB4	0.029	0.026	0.027	0.018	0.017	0.005						
GB2	0.025	0.023	0.023	0.015	0.013	0.011	0.006					
SAB4	0.034	0.030	0.030	0.023	0.020	0.029	0.025	0.006				
SAB1	0.032	0.029	0.029	0.021	0.020	0.018	0.017	0.030	0.007			
GB1	0.032	0.029	0.031	0.024	0.023	0.020	0.019	0.034	0.024	0.011		
GB0	0.036	0.033	0.033	0.027	0.025	0.026	0.023	0.035	0.030	0.025	0.007	
JAB1-NZB2	0.038	0.035	0.034	0.030	0.025	0.028	0.025	0.037	0.031	0.025	0.020	0.004

M2	BA-CC	BA	BA-C	GB3	GB6	GB4	GB2	SAB4	SAB1	GB1	GB0	B1-NZB2
BA-CC	0.006											
BA	0.013	0.010										
BA-C	0.025	0.023	0.000									
GB3	0.018	0.017	0.020	0.010								
GB6	0.022	0.021	0.023	0.012	n/c							
GB4	0.026	0.025	0.026	0.016	0.016	0.003						
GB2	0.026	0.024	0.026	0.015	0.016	0.009	0.009					
SAB4	0.025	0.023	0.028	0.015	0.016	0.020	0.019	0.004				
SAB1	0.025	0.023	0.027	0.014	0.015	0.014	0.014	0.016	0.003			
GB1	0.036	0.037	0.041	0.027	0.027	0.027	0.027	0.030	0.026	0.013		
GB0	0.043	0.043	0.046	0.034	0.033	0.033	0.033	0.036	0.033	0.029	0.007	
JAB1-NZB2	0.042	0.042	0.046	0.033	0.033	0.032	0.032	0.036	0.033	0.028	0.019	0.007

L	BA-CC	BA	BA-C	GB3	GB6	GB4	GB2	SAB4	SAB1	GB1	GB0	B1-NZB2
BA-CC	0.005											
BA	0.010	0.007										
BA-C	0.014	0.012	0.001									
GB3	0.013	0.011	0.011	0.008								
GB6	0.016	0.014	0.014	0.009	n/c							
GB4	0.020	0.019	0.019	0.014	0.014	0.003						
GB2	0.019	0.018	0.017	0.012	0.013	0.007	0.005					
SAB4	0.021	0.019	0.019	0.014	0.014	0.018	0.016	0.004				
SAB1	0.023	0.022	0.021	0.017	0.017	0.014	0.012	0.020	0.004			
GB1	0.026	0.025	0.024	0.020	0.019	0.018	0.015	0.023	0.020	0.008		
GB0	0.029	0.028	0.027	0.023	0.023	0.021	0.018	0.026	0.023	0.017	0.003	
JAB1-NZB2	0.030	0.029	0.029	0.025	0.025	0.023	0.020	0.029	0.025	0.019	0.013	0.006

Figure 4. (**A**) Intergenotypic and intragenotypic p-distance analysis of complete NS1, NS2, N, P, M, SH, G, F, M2, and L genes of RSV-B sequences. (**B**) Unrooted phylogenetic tree of complete NS1, NS2, N, P, M, SH, G, F, M2, and L genes of RSV-B sequences constructed Maximum Likelihood analysis.

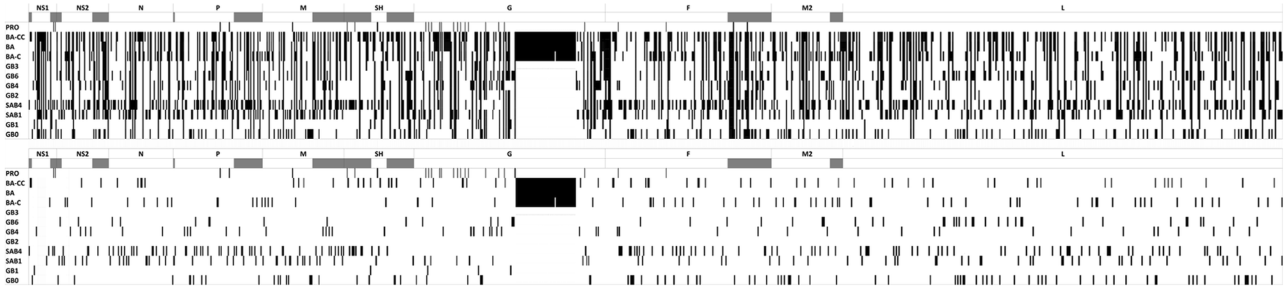


Figure 5. Distribution of molecular markers present in RSV-B genotypes. The location of all (upper panel) and unique (lower panel) molecular markers present in each genotype is shown. Dark bars indicate untranslated regions in which molecular markers were identified.

Around this time, an initial diversification event occurred, with the emergence of JAB1-NZB2, SAB1, URU2, and GB2 genotypes (Figs. 3B and 6). A second diversification event occurred in the mid-90s, leading to the appearance of GB3, GB4, and GB6 genotypes, and characterized by a global spread and predominance of GB3. Of interest, the dataset included five GB3 sequences isolated in the United States obtained from RSV strains isolated between 1996 and 1998 which display the 60-nucleotide duplication described in 2003[33]. These sequences did not group in a single cluster, but were present in four different clades. Furthermore, the sequence of the duplicated region of these early strains showed nucleotide and amino acid differences compared with the initial BA strains described in Buenos Aires, Argentina; nucleotide differences between these GB3 and the original BA viruses were also noted in other genes (NS1, N, P, M, SH, F, M2, and L). After the year 2000, the emergence of CB1-THB occurred; this genotype derived from GB2 and does not have the 60-nucleotide duplication. In addition, during this time the BA-C genotype derived from GB3, containing the partial duplication in the G-gene. These two genotypes have shown an apparent geographic circulation limited to the Asian continent. In contrast,

	Nucleotide molecular markers			Amino acid molecular markers		
Genotype	Unique	Shared	Total	Unique	Shared	Total
BA-CC	71	369	440	10	81	91
BA	0	374	374	0	81	81
BA-C	73	338	411	11	73	84
GB3	0	273	273	0	48	48
GB6	44	280	324	7	53	60
GB4	46	246	292	15	42	57
GB2	0	232	232	0	38	38
SAB4	142	277	419	18	46	64
SAB1	97	226	323	21	47	68
GB1	3	84	87	2	22	24
GB0	71	198	268	8	36	44
JAB1-NZB2	89	176	265	15	32	47
All genotypes	636	577	1,213	107	106	213

Table 1. Number of unique and shared nucleotide and amino acid molecular markers identified in each RSV-B genotype.

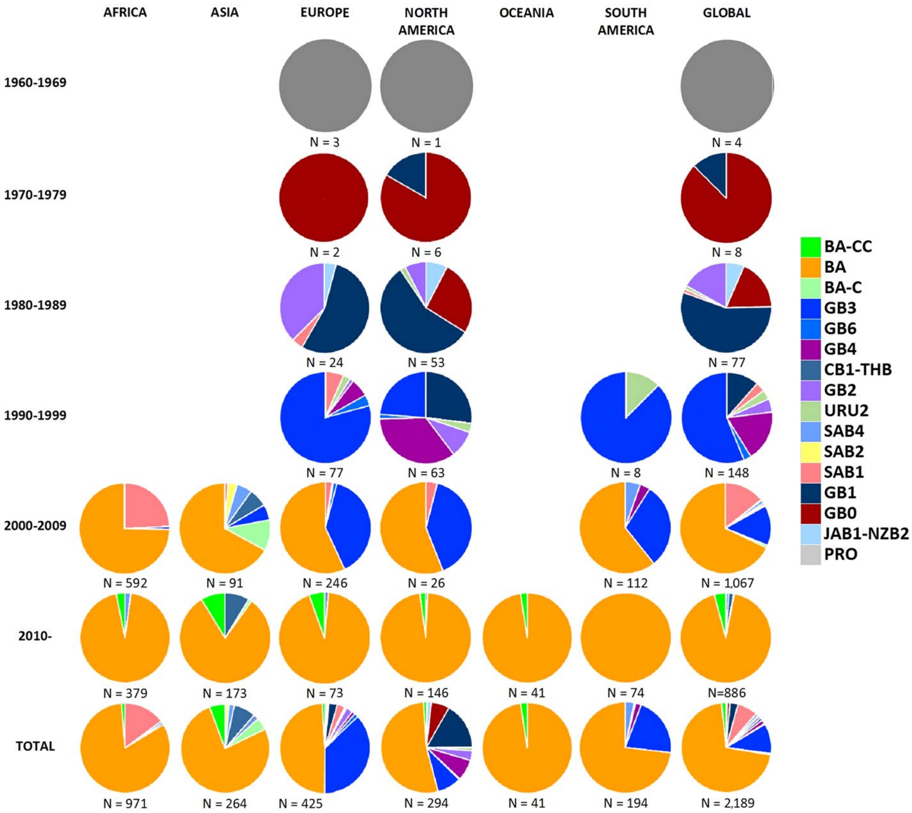


Figure 6. RSV-B genotype distribution since 1960 and each decade thereafter, according to continent of viral detection.

the BA genotype, derived from GB3 sequences with the 60-nucleotide duplication, spread globally and became the predominant genotype worldwide. Finally, the most recent genotype (BA-CC) emerged after 2010 showing a global distribution.

Discussion

Over the last decades, a large number of RSV-B genotypes have been reported in the literature[9–15,23,25,26]. Genotype description is usually based on analysis of a fragment of the G gene of this virus. However, there is no consensus regarding criteria that should be met in order to identify a new genotype, particularly the size of the G gene sequence that needs to be analyzed and the number of sequences that should be included in the analysis[8,20,28,29]. Of relevance, many studies have included a limited number of sequences and have analyzed only the hypervariable region of the G gene. As a result, RSV strains that have been described as distinct genotypes might group in the same cluster when analyzed together[15,25]. These observations, together with the increasing number of sequence data available, underscore the advantage of a systematic approach for viral classification.

Using the approach that we recently described for RSV-A genotype analysis, we analyzed 2,198 sequences of the RSV-B G gene sequence encoding for the complete G protein ectodomain and found that when all available sequences (as of September, 2018) were analyzed, many of previously described genotypes cluster with other strains previously considered as unique genotypes. As such, the number of distinct RSV-B genotypes identified through this analysis was fifteen (in contrast to at least 37 genotypes mentioned in the literature). Analysis of full sequences of each of the complete genes from 478 RSV-B strains for which the complete genome sequence was available supported the inclusion of many previously described genotypes within a smaller number of genotypes. It is noteworthy that although analysis of other genes did not allow for distinction between all genotypes in all comparisons, in many instances this result can be explained by the paucity of sequences available for some RSV-B genotypes. For instance, there was only one full genome sequence available for the GB6 genotype, with comparisons that include this genotype accounting for 20 (40%) of the 50 instances in which intergenotypic p-distance did not support classification of a specific clade as a distinct genotype. Overall, 30 (10.9%) of the 275 intergenotypic comparisons that included 20 sequences or less did not allow to differentiate between genotypes in contrast to 4.9% of comparisons that included more than 20 sequences. This highlights that inclusion of a minimum number of sequences of each genotype might be required for definite genotype assignment.

One of the most notable distinct features of some RSV-B genotypes is the presence of a partial duplication of the G gene, initially described in Buenos Aires, Argentina, and termed BA strains[33]. Over the past two decades, RSV-B strains harboring this partial duplication have become the predominant RSV-B viruses. As a result of global expansion and diversification, a large number of BA genotypes have been described. Intra- and intergenotypic comparisons between BA genotypes indicated that these strains can be classified within four distinct groups: early BA strains (which clustered within the GB3 genotype), late BA strains, BA-C, and BA-CC genotypes. While the partial duplication of the G gene is the landmark characteristic of BA strains, the p-distance between early BA strains and GB3 strains was below the threshold to consider them distinct genotypes. This result is similar to analysis of RSV-A strains with a partial duplication of the G gene (ON1 strains) which also present a low p-distance compared to RSV-A viruses without the duplication (NA1 genotype) and, as a result, have been considered as part of the NA1 genotype[8,30]. In addition, five GB3 sequences isolated between 1996 and 1998 displayed the 60-nucleotide partial duplication of the G gene. These sequences did not group in a single cluster, but were present in four different clades. In addition, these GB3 strains showed nucleotide and amino acid differences compared with the initial BA strains described in Argentina in almost all genes. These observations suggest the occurrence of independent duplication events that, ultimately, resulted in the establishment of a dominant variant leading to the emergence of the BA genotype. This is consistent with previous reports that indicate that more than one duplication event resulted in new variants of human metapneumovirus and RSV-A[34–36].

The definition of an early cluster of RSV-B strains as a distinct genotype (which we have termed GB0) is supported by the phylogenetic analysis, as well as the G gene intergenotypic p-distance analysis. The p-distance between the proposed early genotype and all other genotypes was higher than the value established as a threshold to identify a distinct genotype. In addition, analysis of all other RSV genes (except NS1 and N, for which there were two and one exceptions, respectively) supported the identification of these strains as a distinct genotype. Also, we identified 71 and 8 nucleotide and amino acid markers, respectively that are distinct for this genotype. These markers were found in 10 genes and 5 deduced proteins. Circulation of this genotype occurred in North America and Europe between 1972 and 1983.

Distinct molecular markers have been previously described for several RSV-B genotypes. For instance, BA13 had been reported to display unique amino acid changes (T232A, K233G, T240K/G, R242G, Q248R, D253G, T255A, T256A, K258G, D263Y and E292K)[12]. However, based on analysis of a large sequence dataset, we observed that many of these markers were not exclusive of BA13; for instance, R242G was also found to be present in GB6, and Q248R in GB4. Another example is BA9, which had been described as having two specific clusters named ATI and TRT based on substitutions at positions 107, 136, and 254 of the G protein (A107, T107, T136, R136, T254 and I254) and at positions 173 and 209 of the F protein (S173, L173, K209 and Q209)[37]; however, we observed that T107A and S172L are markers for BA-CC genotype and R136I for URU2. It is noteworthy that G protein amino acid substitutions T107A and T254I (markers of BA-CC), R136I (marker of URU2), and K258D/N (markers of GB4 and CB1-THB, respectively) may alter O-glycosylation patterns and, as a result, may affect antigenicity and facilitate homologous reinfections[37]. Nine unique amino acid markers were located at the F protein. However, none was found at the antigenic site targeted by palivizumab; in fact, all RSV-B F gene sequences included in the analysis were conserved at antigenic site II (aa 255–275). Five sequences (1.03% of the dataset), all corresponding to genotype BA that circulated between 2012–2014, showed the S276N substitution. Overall, we identified 1,213 nucleotide and 213 amino acid molecular markers. As previously noted, GB0 strains displayed eight unique amino acid molecular markers. In addition, this proposed early genotype had 185 nucleotide and 36 amino acid molecular markers which were shared with one or more genotypes.

Overall, our analysis allowed to identify molecular markers that at this time can be considered as specific of certain genotypes, particularly when several of them are identified together. While we identified certain markers with high specificity for a single genotype, many markers are shared by two or more genotypes. Therefore, it is likely that availability of more sequencing information, particularly from contemporaneous or future RSV strains, might modify the specificity estimations obtained by us. On the other hand, ongoing monitoring of the prevalence of these markers on currently circulating genotypes might help identity the emergence of new genotypes in the future.

Analysis of the temporal and geographical distribution of the different RSV-B genotypes showed that after the report of this RSV subgroup in 1960, two distinct genotypes circulated in North America and Europe, GB1 and the novel early genotype termed G0 in this report (Fig. 6). During the decade between 1980 and 1989 viral strains belonging to these genotypes continued to be the predominant RSV-B viruses, with the appearance of additional genotypes, namely JAB1-NZB2, GB2, SAB1, and URU2. Between 1990 and 1999 RSV-B strains displayed further diversification including GB3, GB4, and GB6; of note, viral strains with a partial duplication of the G gene emerged within the GB3 genotype (described as BA genotype by Trento et al.[33]) during this time. After the year 2000, further diversification of RSV-B strains led to the development of new genotypes with (BA, BA-C, and BA-CC genotypes) and without (CB1-THB) the partial duplication of the G gene. Interestingly, the temporal evolution of RSV-A strains has shown a similar pattern, although there have been fewer genotypes described, and the emergence of RSV-A strains with a partial duplication of the G gene (ON1 strains), analogous to the BA strains, occurred approximately ten years later[36,38]. This could be explained, in part, by the fact that RSV-B evolution rate is higher than that of RSV-A.

In summary, despite displaying a wide diversity, RSV-B strains can be grouped in 15 distinct genotypes. The 60-nt partial duplication of the G gene does not identify a unique genotype and includes viral strains within four different genotypes (GB3, BA, BA-C, and BA-CC). Finally, we have identified a previously unrecognized early genotype which we have termed as GB0, since it circulated prior to the emergence of the GB1 genotype.

Materials and methods

Dataset selection and curation.

The dataset for this study included all RSV-B strains for which at least the complete G gene ectodomain sequence had been deposited in GenBank up until September 2018. All RSV sequences available on NCBI were downloaded and, as this study focused on the analysis of the complete ectodomain of circulating RSV strains, several inclusion criteria had to be met in order to proceed with subsequent analyses. Synthetic RSV sequences, sequences from organisms other than RSV, and sequences with nucleotide length smaller than the complete ectodomain length were excluded. 10,340 sequences fulfilled the inclusion criteria and were downloaded.

Blast2GO v5.2.5 software was used to analyze the sequences and identify the strains corresponding to RSV-B; local BLAST was carried out against a database of 10 RSV-B G gene ectodomain reference sequences resulting in 3,029 RSV-B G gene sequences[39]. These sequences were aligned using MAFFT v7.450 and, subsequently, manually inspected and aligned if needed using BioEdit v7.0.5.3[40,41]. The sequences were inspected and those that presented gaps (other than the partial 60 nt duplication described in 2000)[33], degenerate nucleotides, insertions or deletions at the initial, middle or terminal ectodomain (with exception of a six nucleotide deletion present at position 473–478 in 27% of the sequences) were eliminated from the dataset to prevent genotype misassignment. 2,198 RSV-B complete G ectodomain sequences fulfilled the criteria previously described and, for each sequence, information such as accession number, strain, year and country of isolation was registered.

Two different alignments were obtained from this dataset: the first included only the second hypervariable region (spanning from nt 645 to the end of the G gene); the second alignment included the complete G gene ectodomain (spanning from nt 312 to the end of the G gene).

The complete dataset consisted of 2198 complete G gene ectodomain sequences, of which 18 sequences consisted of the complete G gene, 29 consisted of the complete G and SH genes, and 478 consisted of complete genome sequences; the 10 RSV genes were trimmed from 3′UTR to 5′UTR and aligned using MUSCLE algorithm and duplicated sequences were removed from each gene alignment to perform the cladistic analyses[42].

Reference sequences selection.

Reference sequences were selected through an extensive search of the literature. First, we identified 736 sequences included in 22 articles published between 2003 to 2018 as genotype references[5,6,10,13–16,23–26,33,43–52]; subsequently, these were assessed to verify concordance in genotype assignment; some sequences could not be identified or traced due to assignment of IDs different from strain names or GenBank accession numbers. A total of 691 of these 736 sequences were traceable. Sequences that had been used by two or more of the 22 authors as reference sequences and those that have been identified as unique genotypes were selected, resulting in a total of 188 sequences (Supplementary Table 2). Sequences which had been assigned by two or more authors as representative of a different genotype were discarded (sequences with discordant genotype identity). In the case of recently or uniquely identified genotypes the first sequences submitted to NCBI were used as reference sequences. Genotype assignment agreed between two or more authors in 115 of 188 sequences; in addition, 64 sequences from recently or uniquely identified genotypes were included as references based only on their original description[5–7,10,13–16,18,19,23–26,33,44–50,52]. The length of 122 of the resulting 188 reference sequences was shorter than the complete G gene ectodomain. To resolve this limitation during genotype assignment of sequences included in the study dataset, we selected equivalent references using a Maximum Likelihood analysis of the G gene second hypervariable region under the GTR + Γ + I model and 1,000 bootstrap iterations, as previously described[8]. For all subsequent analyses, 169 original or equivalent reference sequences listed in Supplementary Table 3 were used.

Genotype assignment.

Topologies from Maximum Likelihood analysis under GTR + Γ + I model with 1,000 bootstrapping iterations inferred with MEGA X v10.0.3, as well as a Maximum Clade Credibility Tree generated from a Bayesian Skyline Plot Analysis inferred with BEAST v2.5.1 package were visualized on FigTree v1.4 and considered for genotype assignment by clade clustering with reference or equivalent sequences[53]. P-distance matrices were generated with MEGA X v10.0.3 to calculate intragenotypic and intergenotypic distance for the G gene ectodomain as well as the 10 complete genes sequences.

The Maximum Clade Credibility Tree was also used to recreate the history of major changes of RSV-B over time and was generated from unique complete G gene ectodomain sequences using TreeAnnotator v2.5.1 from the corresponding phylogenetic analysis by the MCMC method performed with BEAST v2.5.1 package. Bayesian Skyline method was used to analyze the dataset assuming both relaxed and strict molecular clock. MCMC were run 400,000,000 steps and sampled every 20,000 steps; convergence achievement was confirmed with Tracer v1.7.1.

All 10 RSV-B genes (NS1, NS2, N, P, M, SH, G, F, M2, and L) were characterized and analyzed via cladistic analysis and calculation of intragenotypic and intergenotypic p-distance matrices using MEGA X v10.0.3 software.

In addition, we carried out recombination analyses with the use of RDP4 v.4.100 to identify potential recombinants; the following algorithms were included in the analysis: RDP, GENECONV, BOOTSCAN, MaxChi, CHIMAERA, SISCAN, and 3SEQ[31]. In addition, the GARD algorithm was also used to confirm the presence of recombination events[32].

Molecular markers detection. To detect molecular markers distributed all along the 10 RSV-B genes, sequences were grouped according to the previously assigned genotype, sequences were aligned spanning from 3′UTR to 5′UTR and translated using BioEdit v7.0.5.3. Differences of each grouped genotype with respect to RSV-B reference sequence "strain B1" (accession number NC_001781.1) were recorded[54]; the differences were considered molecular markers if a nucleotide or amino acid shift occurred at a site with respect to the reference sequence in 75% or more of the genotyped sequences. This analysis included 478 complete genome sequences for 8 of 10 genes (NS1, NS2, N, P, M, F, M2, and L), 507 sequences for SH, and 525 for complete G gene.

Data availability

This study was carried out with data retrieved from GenBank. All data used is available in public databases.

Received: 30 August 2020; Accepted: 22 January 2021
Published online: 10 February 2021

References

1. Griffiths, C., Drews, S. & Marchant, D. Respiratory syncytial virus : Infection, detection, and new options for prevention and treatment. *Clin. Microbiol. Rev.* **30**, 277–319 (2017).
2. Hendry, M. R. *et al.* Concurrent circulation of antigenically distinct strains of respiratory syncytial virus during community outbreaks. *J. Infect. Dis.* **153**, 291–297 (1986).
3. Mufson, M. A., Orvell, C., Rafnar, B. & Norrby, E. Two distinct subtypes of human respiratory syncytial virus. *J. Gen. Virol.* **66**, 2111–2124 (1985).
4. Peret, T. C. T. *et al.* Circulation patterns of group A and B human respiratory syncytial virus genotypes in 5 communities in North America. *J. Infect. Dis.* **181**, 1891–1896 (2000).
5. Ahmed, A. *et al.* Co-circulation of 72bp duplication group A and 60bp duplication group B respiratory syncytial virus (RSV) strains in Riyadh, Saudi Arabia during 2014. *PLoS ONE* **11**, e0166145 (2016).
6. Trento, A. *et al.* Ten years of global evolution of the human respiratory syncytial virus BA genotype with a 60-nucleotide duplication in the G protein gene. *J. Virol.* **84**, 7500–7512 (2010).
7. Zhang, R.-F. *et al.* Human respiratory syncytial virus in children with acute respiratory tract infections in China. *J. Clin. Microbiol.* **48**, 4193–4199 (2010).
8. Muñoz-Escalante, J. C. *et al.* Respiratory syncytial virus A genotype classification based on systematic intergenotypic and intragenotypic sequence analysis. *Sci. Rep.* **9**, 20097 (2019).
9. Ren, L., Xiao, Q., Zhou, L., Xia, Q. & Liu, E. Molecular characterization of human respiratory syncytial virus subtype B: A novel genotype of subtype B circulating in China. *J. Med. Virol.* **87**, 1–9 (2015).
10. Gimferrer, L. *et al.* Circulation of a novel human respiratory syncytial virus Group B genotype during the 2014–2015 season in Catalonia (Spain). *Clin. Microbiol. Infect.* **22**(97), e5-97.e8 (2015).
11. Bashir, U. *et al.* Molecular detection and characterization of respiratory syncytial virus B genotypes circulating in Pakistani children. *Infect. Genet. Evol.* **47**, 125–131 (2017).
12. Aamir, U. B. *et al.* Molecular characterization of circulating respiratory syncytial virus genotypes in Pakistani children, 2010–2013. *J. Infect. Public Health* **13**, 438–445 (2020).
13. Gaymard, A. *et al.* Genetic characterization of respiratory syncytial virus highlights a new BA genotype and emergence of the ON1 genotype in Lyon, France, between 2010 and 2014. *J. Clin. Virol.* **102**, 12–18 (2018).
14. Ábrego, L. E. *et al.* Genetic variability of human respiratory syncytial virus group B in Panama reveals a novel genotype BA14. *J. Med. Virol.* **89**, 1734–1742 (2017).
15. Cui, G. *et al.* Genetic variation in attachment glycoprotein genes of human respiratory syncytial virus subgroups A and B in children in recent five consecutive years. *PLoS ONE* **8**, e75020 (2013).
16. Choudhary, M. L., Anand, S. P., Wadhwa, B. S. & Chadha, M. S. Genetic variability of human respiratory syncytial virus in Pune Western India. *Infect. Genet. Evol.* **20**, 369–377 (2013).
17. Houspie, L. *et al.* Circulation of HRSV in Belgium: From multiple genotype circulation to prolonged circulation of predominant genotypes. *PLoS ONE* **8**, e60416 (2013).
18. Kuroiwa, Y. *et al.* A phylogenetic study of human respiratory syncytial viruses group A and B strains isolated in two cities in Japan from 1980–2002. *J. Med. Virol.* **76**, 241–247 (2005).
19. Matheson, J. W. *et al.* Distinct patterns of evolution between respiratory syncytial virus subgroups A and B From New Zealand isolates collected over thirty-seven years. *J. Med. Virol.* **78**, 1354–1364 (2006).
20. Venter, M., Madhi, S. A., Tiemessen, C. T. & Schoub, B. D. Genetic diversity and molecular epidemiology of respiratory syncytial virus over four consecutive seasons in South Africa: Identification of new subgroup A and B genotypes. *J. Gen. Virol.* **82**, 2117–2124 (2001).

21. Blanc, A., Delfraro, A., Frabasile, S. & Arbiza, J. Genotypes of respiratory syncytial virus group B identified in Uruguay. *Arch. Virol.* **150**, 603–609 (2005).
22. Trento, A. *et al.* Natural history of human respiratory syncytial virus inferred from phylogenetic analysis of the attachment (G) glycoprotein with a 60-nucleotide duplication. *J. Virol.* **80**, 975–984 (2006).
23. Dapat, I. C. *et al.* New genotypes within respiratory syncytial virus group B genotype BA in Niigata Japan. *J. Clin. Microbiol.* **48**, 3423–3427 (2010).
24. Baek, Y. H. *et al.* Prevalence and genetic characterization of respiratory syncytial virus (RSV) in hospitalized children in Korea. *Arch. Virol.* **157**, 1039–1050 (2012).
25. Auksornkitti, V. *et al.* Molecular characterization of human respiratory syncytial virus, 2010–2011: Identification of genotype ON1 and a new subgroup B genotype in Thailand. *Arch. Virol.* **159**, 499–507 (2014).
26. Arnott, A. *et al.* A study of the genetic variability of human respiratory syncytial virus (HRSV) in Cambodia reveals the existence of a new HRSV group B genotype. *J. Clin. Microbiol.* **49**, 3504–3513 (2011).
27. Khor, C. S., Sam, I. C., Hooi, P. S. & Chan, Y. F. Displacement of predominant respiratory syncytial virus genotypes in Malaysia between 1989 and 2011. *Infect. Genet. Evol.* **14**, 357–360 (2013).
28. Agoti, C. N. *et al.* Successive respiratory syncytial virus epidemics in local populations arise from multiple variant introductions, providing insights into virus persistence. *J. Virol.* **89**, 11630–11642 (2015).
29. Goya, S. *et al.* Toward unified molecular surveillance of RSV: A proposal for genotype definition. *Influenza Other Respir. Viruses* **14**, 274–285 (2020).
30. Trento, A. *et al.* Conservation of G-protein epitopes in respiratory syncytial Virus (Group A) Despite broad genetic diversity: Is antibody selection involved in virus evolution?. *J. Virol.* **89**, 7776–7785 (2015).
31. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **1**, 1–5 (2015).
32. Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. W. GARD: A genetic algorithm for recombination detection. *Bioinformatics* **22**, 3096–3098 (2006).
33. Trento, A. *et al.* Major changes in the G protein of human respiratory syncytial virus isolates introduced by a duplication of 60 nucleotides. *J. Gen. Virol.* **84**, 3115–3120 (2003).
34. Saikusa, M. *et al.* A novel 111-nucleotide duplication in the G gene of human metapneumovirus. *Microbiol. Immunol.* **61**, 507–512 (2017).
35. Piñana, M. *et al.* Insights into immune evasion of human metapneumovirus: Novel 180- and 111-nucleotide duplications within viral G gene throughout 2014–2017 seasons in Barcelona, Spain. *J. Clin. Virol.* **132**, 2 (2020).
36. Comas-García, A., Noyola, D. E., Cadena-Mota, S., Rico-Hernández, M. & Bernal-Silva, S. Respiratory syncytial virus-A ON1 genotype emergence in central Mexico in 2009 and evidence of multiple duplication events. *J. Infect. Dis.* **217**, 1089–1098 (2018).
37. Okamoto, M. *et al.* Molecular characterization of respiratory syncytial virus in children with repeated infections with subgroup B in the philippines. *J. Infect. Dis.* **218**, 1045–1053 (2018).
38. Eshaghi, A. *et al.* Genetic variability of human respiratory syncytial virus a strains circulating in Ontario: A novel genotype with a 72 nucleotide G gene duplication. *PLoS ONE* **7**, e32807 (2012).
39. Götz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435 (2008).
40. Hall, T. BioEdit: A user friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucleic Acids Symp. Ser.* **4**, 95–98 (1999).
41. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
42. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**, 636–641 (2019).
43. Zhang, T. *et al.* Tracing the emerging genotypes of human respiratory syncytial virus in Beijing by evolution analysis of the attachment glycoprotein (G) gene. *Infect. Genet. Evol.* **65**, 18–27 (2018).
44. Yu, X. *et al.* Human respiratory syncytial virus in children with lower respiratory tract infections or influenza-like illness and its co-infection characteristics with viruses and atypical bacteria in Hangzhou China. *J. Clin. Virol.* **69**, 1–6 (2015).
45. Tabatabai, J., Prifert, C., Pfeil, J., Grulich-Henn, J. & Schnitzler, P. Novel respiratory syncytial virus (RSV) genotype ON1 predominates in germany during winter season 2012–13. *PLoS ONE* **9**, e109191 (2014).
46. Almajhdi, F. N., Farrag, M. A. & Amer, H. M. Group B strains of human respiratory syncytial virus in Saudi Arabia: Molecular and phylogenetic analysis. *Virus Genes* **48**, 252–259 (2014).
47. Zlateva, K. T., Lemey, P., Moës, E., Vandamme, A.-M. & Van Ranst, M. Genetic variability and molecular evolution of the human respiratory syncytial virus subgroup B attachment G protein. *J. Virol.* **79**, 9157–9167 (2005).
48. Araújo Moura, F. E. *et al.* Genetic diversity of respiratory syncytial virus isolated during an epidemic period from children of Northeastern Brazil. *J. Med. Virol.* **74**, 156–160 (2004).
49. Haider, M. S. H. *et al.* BA9 lineage of respiratory syncytial virus from across the globe and its evolutionary dynamics. *PLoS ONE* **13**, e0193525 (2018).
50. Song, J. *et al.* Emergence of BA9 genotype of human respiratory syncytial virus subgroup B in China from 2006 to 2014. *Sci. Rep.* **7**, 16765 (2017).
51. Fan, R. *et al.* Respiratory syncytial virus subtype ON1/NA1/BA9 predominates in hospitalized children with lower respiratory tract infections. *J. Med. Virol.* **89**, 213–221 (2017).
52. Gardinassi, L. *et al.* Diversity and adaptation of human respiratory syncytial virus genotypes circulating in two distinct communities: Public hospital and day care center. *Viruses* **4**, 2432–2447 (2012).
53. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15**, e1006650 (2019).
54. Karron, R. A. *et al.* Respiratory syncytial virus (RSV) SH and G proteins are not essential for viral replication in vitro: Clinical evaluation and molecular characterization of a cold-passaged, attenuated RSV subgroup B mutant. *Proc. Natl. Acad. Sci.* **94**, 13961–13966 (1997).

Acknowledgements

This study was supported by CONACYT Convocatoria de Investigación Científica Básica (Proyecto CONACYT-A1-S-38080).

Author contributions

J.C.M.E. and D.E.N. contributed to the conception and design of the study; acquisition, analysis, and interpretation of data; drafting of the manuscript; and approval of the final version of the manuscript; A.C.G., contributed to the conception and design of the study; acquisition, analysis, and interpretation of data; revision the manuscript; and approval of the final version of the manuscript; S.B.S. contributed to analysis and interpretation of data; revision the manuscript; and approval of the final version of the manuscript.

Competing interests

JCME, ACG, and SBS declare that there are no competing interests regarding the publication of this manuscript. DEN has participated as a member of the speakers' bureau of AbbVie and speakers' bureau and advisory board for Sanofi Pasteur.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-83079-2.

Correspondence and requests for materials should be addressed to D.E.N.