



## ORIGINAL ARTICLE

# Toward unified molecular surveillance of RSV: A proposal for genotype definition

Stephanie Goya<sup>1,2</sup>  | Mónica Galiano<sup>3</sup> | Inne Nauwelaers<sup>4</sup> | Alfonsina Trento<sup>5</sup> | Peter J. Openshaw<sup>4</sup> | Alicia S. Mistchenko<sup>1,6</sup> | Maria Zambon<sup>3</sup> | Mariana Viegas<sup>1,2</sup> 

<sup>1</sup>Virology Laboratory, Ricardo Gutiérrez Children's Hospital, Buenos Aires, Argentina

<sup>2</sup>Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET), Buenos Aires, Argentina

<sup>3</sup>Respiratory Virus Unit, National Infection Services, Public Health England, London, UK

<sup>4</sup>National Heart and Lung Institute, Imperial College London, London, UK

<sup>5</sup>Instituto de Investigación Hospital 12 de Octubre, Madrid, Spain

<sup>6</sup>Comisión de Investigaciones Científicas (CIC), Buenos Aires, Argentina

## Correspondence

Stephanie Goya and Mariana Viegas,  
Virology Laboratory, Ricardo Gutiérrez  
Children's Hospital, Gallo 1330, C1425EFD  
Buenos Aires, Argentina.  
Emails: goyastephanie@gmail.com (SG);  
viegasmariana@conicet.gov.ar (MV)

## Present address

Mónica Galiano, WHO Collaborating Centre  
for Reference and Research on Influenza,  
Francis Crick Institute, London, UK

## Funding information

This research did not receive specific funding from agencies in the public, commercial, or not-for-profit sectors. Consejo Nacional de Investigaciones Científicas y Técnicas supported SG and MV. Comisión de Investigaciones Científicas de la Provincia de Buenos Aires supported ASM. IN, MG, PO and MZ are supported by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Respiratory Infections and the Biomedical Research Centre (BRC) at Imperial College Healthcare NHS Trust, funded by NIHR in partnership with PHE. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

## Abstract

**Background:** Human respiratory syncytial virus (RSV) is classified into antigenic subgroups A and B. Thirteen genotypes have been defined for RSV-A and 20 for RSV-B, without any consensus on genotype definition.

**Methods:** We evaluated clustering of RSV sequences published in GenBank until February 2018 to define genotypes by using maximum likelihood and Bayesian phylogenetic analyses and average p-distances.

**Results:** We compared the patterns of sequence clustering of complete genomes; the three surface glycoproteins genes (SH, G, and F, single and concatenated); the ectodomain and the 2nd hypervariable region of G gene. Although complete genome analysis achieved the best resolution, the F, G, and G-ectodomain phylogenies showed similar topologies with statistical support comparable to complete genome. Based on the widespread geographic representation and large number of available G-ectodomain sequences, this region was chosen as the minimum region suitable for RSV genotyping. A genotype was defined as a monophyletic cluster of sequences with high statistical support ( $\geq 80\%$  bootstrap and  $\geq 0.8$  posterior probability), with an intragenotype p-distance  $\leq 0.03$  for both subgroups and an intergenotype p-distance  $\geq 0.09$  for RSV-A and  $\geq 0.05$  for RSV-B. In this work, the number of genotypes was reduced from 13 to three for RSV-A (GA1-GA3) and from 20 to seven for RSV-B (GB1-GB7). Within these, two additional levels of classification were defined: subgenotypes and lineages. Signature amino acid substitutions to complement this classification were also identified.

The peer review history for this article is available at <https://publons.com/publon/10.1111/irv.12715>

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Influenza and Other Respiratory Viruses* published by John Wiley & Sons Ltd

**Conclusions:** We propose an objective protocol for RSV genotyping suitable for adoption as an international standard to support the global expansion of RSV molecular surveillance.

**KEYWORDS**

average genetic distance, genotypes, global molecular surveillance, human orthopneumovirus, human respiratory syncytial virus, lineages, phylogenetic analysis, subgenotypes

## 1 | BACKGROUND

Human respiratory syncytial virus (RSV) is the commonest viral cause of acute lower respiratory tract infections in children worldwide, being the main infectious reason for pediatric hospitalizations.<sup>1</sup> There is yet neither an effective vaccine nor antiviral therapy, and treatment remains supportive, although passive antibody prophylaxis (palivizumab) is available in developed countries for prevention of high-risk babies. Progress is being made with vaccine development with encouraging results reported for live-attenuated and subunit approaches.<sup>2</sup>

RSV is member of the *Orthopneumovirus* genus within the family *Pneumoviridae*.<sup>3</sup> It is an enveloped virus with a negative-sense ssRNA genome of ~15 200 nucleotides (nt) in length. Its genome encodes for 11 proteins (Figure 1A). Two antigenic subgroups (A and B) are distinguished by polyclonal and monoclonal antibodies. Both subgroups are evolutionary lineages which diverged approximately 350 years ago<sup>4</sup> with considerable genotypic variability within them. The major differences are found in the attachment glycoprotein G, which has only 53% amino acid sequence conservation across strains and has been used historically for molecular characterization.<sup>5</sup>

Currently, 13 RSV genotypes have been defined among the subgroup A strains (GA1-7,<sup>6,7</sup> SAA1,<sup>8</sup> NA1-4,<sup>9,10</sup> and ON1-2<sup>11,12</sup>) and 20 genotypes for the subgroup B strains (GB1-4,<sup>6</sup> SAB1-4,<sup>8</sup> URU1-2,<sup>13</sup> and BA1-10<sup>14,15</sup>) but the criteria used for definition of a genotype varies based on phylogenetic analyses inferred using different methods (maximum likelihood, maximum parsimony, neighbor-joining or Bayesian inferences). Most definitions focus on clustering of phylogenetic clades with significant bootstrap values (>70%) in trees built from alignments encompassing the 2nd hypervariable region (HR) of G gene (~270 nt in length). To support these definitions, the average genetic distance (p-distance) has been used, sometimes as an informative tool, sometimes as an arbitrarily selected cut-off value (<0.07)<sup>8</sup> or similarity value (>96%).<sup>6</sup> Overall, there is a lack of consensus regarding the criteria to be used to allocate genotypes. The presence of a duplicated segment (ON -72 nt duplication- and BA -60 nt duplication- genotypes in RSV-A and RSV-B, respectively) in the 2nd HR of G gene has been used as an added criterion to define new genotypes.<sup>11,16</sup> A more recent proposal for genotyping RSV-A strains used phylogenetic analysis of the G-ectodomain and reevaluated historical genotypes using average p-distances within and among genotypes with

a cut-off value of 0.049, based on the average p-distance of the oldest RSV-A genotype, GA1.<sup>17</sup>

Unification of the nomenclature and phylogenetic classification of viruses with high impact in human and animal health, such as highly pathogenic H5N1 avian influenza virus ([https://www.who.int/influenza/gisrs\\_laboratory/h5n1\\_nomenclature/en/](https://www.who.int/influenza/gisrs_laboratory/h5n1_nomenclature/en/)), Newcastle disease virus,<sup>18</sup> measles virus,<sup>19</sup> and HCV<sup>20</sup> is an important underpinning principle for unambiguous tracking of virus evolution, which may have significant public health consequences. Reaching consensus on a unified criterion for RSV genotype definition is essential to explore the association of genotype with disease severity, or geographic or temporal restriction of virus circulation. The aim of this work is to reach a new genotype definition based on both phylogenetic analyses and average p-distances that would bring uniformity to strains designations and thereby facilitate conclusions about viral evolution based on data from surveillance studies.

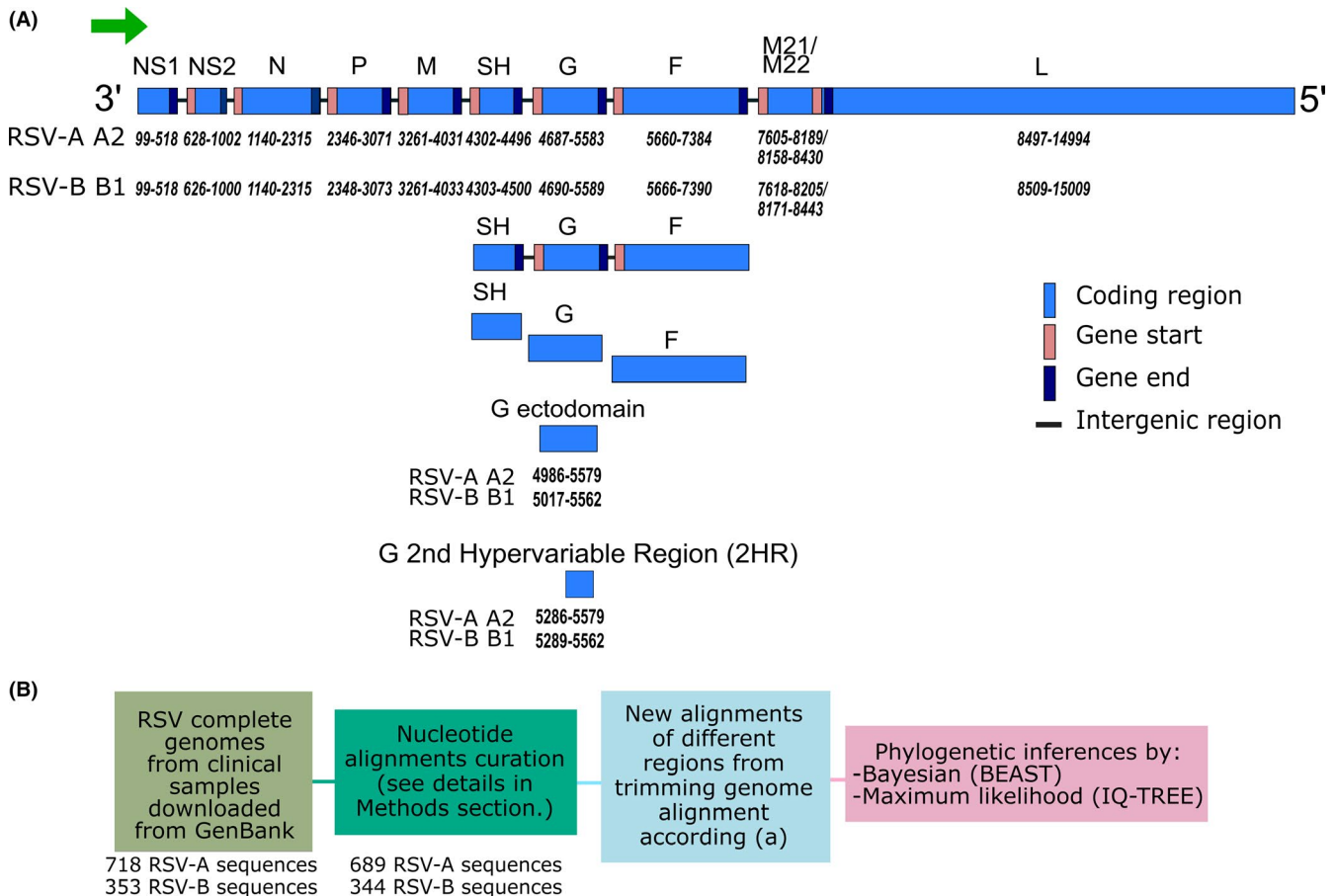
## 2 | METHODS

### 2.1 | Sequence datasets

RSV complete genome sequences from human clinical samples were retrieved from GenBank up to February 2018 (718 RSV-A and 348 RSV-B sequences). The criteria used to curate the sequences include removing sequences with “NNN” regions, ambiguous nucleotides, and/or sequences with 1-2 nucleotide deletions causing frameshifts. Sequences with incorrect RSV subgroup allocation were identified and added to the correct subgroup. After curation, a total of 689 complete genomes of RSV-A and 344 of RSV-B sequences were obtained (Figure 1B).

Each genomic dataset was aligned with MUSCLE (multiple sequence comparison by log-expectation).<sup>21</sup> These were further trimmed to produce alternative datasets encompassing different regions of the RSV genome to be analyzed: the concatenated SH-G-F genes (including their intergenic regions) as a single stretch; the three surface glycoprotein genes as separate ORFs, the ectodomain and the 2nd HR of the G gene (Figure 1A).

A second analysis was performed with all the G-ectodomain sequences published in GenBank up to February 2018 (3362 RSV-A and 1742 RSV-B sequences). Alignment curation was performed as described for genome sequences. In addition, identical nucleotide



**FIGURE 1** Scheme of RSV genome organization and workflow for different RSV alignments and phylogenies generation. A, RSV genome organization and the regions used for new alignments generated from trimming of the genomes. Numbers below the rectangles show the position of the coding regions in the genome and allocation of G-ectodomain and G 2nd hypervariable regions, according to the reference strains A2 for RSV-A (GenBank acc. no.:NC\_038235.1) and B1 for RSV-B (GenBank acc. no.: NC\_001781.1). B, Workflow for the phylogenies inference of different regions used for both RSV-A and B. Numbers of sequences before and after data curation are detailed

sequences were detected and removed with IQ-TREE software and only one representative non-identical sequence was kept.<sup>22</sup> After curation, a total of 2481 G-ectodomain sequences for RSV-A and 1259 for RSV-B were obtained.

The final alignments were visually checked for artifacts produced during the alignment procedure.

## 2.2 | Phylogenetic inferences

The selection of the most suitable nucleotide substitution models was performed with IQ-TREE software according to the Akaike information criterion (AIC).<sup>23</sup> The GTR + G was the most suitable model for most of the alignments, with exception of complete genome and the three surface glycoprotein alignments in which the GTR + I + G was selected for both subgroups. In addition, TIM + G was the model selected for both SH alignments.

Maximum likelihood trees were inferred using IQ-TREE v1.6.7 software with 1000 ultrafast bootstrap replicates plus SH-like approximate likelihood ratio test as statistical support (values  $\geq 80\%$  were defined as well-supported).<sup>24</sup>

Bayesian trees were inferred with BEAST v1.10.2 package.<sup>25</sup> Demographic and molecular clock model selections were performed by estimating the model marginal log-likelihood through the path sampling method, and uncorrelated relaxed molecular clock and Skyride tree prior were selected. The number of generations was between 50 and 100 million, and an appropriate sample frequency was used to obtain 10 000 trees. Convergence was assessed by estimating the effective sampling size after a 10% of burn-in by using TRACER v1.7.1 (<http://tree.bio.ed.ac.uk/software/tracer/>). TreeAnnotator was used to summarize the posterior trees from BEAST into a maximum clade credibility tree (MCCT). The statistical support of the nodes was considered as well-supported when posterior probabilities were  $\geq 0.8$ .

A web tool ([www.phylo.io](http://www.phylo.io)) was used for comparison of tree topologies.<sup>26</sup> FIGTREE v1.4.3 software was used for visualization of phylogenies.<sup>27</sup>

Throughout the entire manuscript, we use the term genetic clade to mean a monophyletic cluster with high statistical support ( $\geq 80\%$  bootstrap and  $\geq 0.8$  posterior probability values) with more than three epidemiologically unrelated sequences, that is, non-identical sequences from different countries, different outbreaks, or combination of both.

### 2.3 | Evaluation of phylogenetic signal

The loss of phylogenetic signal due to substitution saturation was evaluated with DAMBE software. The level of saturation was studied by plotting the pairwise number of observed transitions and transversions versus genetic distance.<sup>28</sup>

### 2.4 | Genetic distances and amino acids analyses

The average genetic distance within and among clades was estimated for alignments with unique sequences with MEGA7 software with the most simplified method, p-distance, as a proportion of nucleotide sites at which two sequences being compared were different.<sup>29</sup> Pairwise deletion was the treatment used for the alignment's gaps. The standard errors of the estimates were determined by the bootstrap method with 1000 replicates.

The TREESUB program (<https://github.com/tamuri/treesub>) was used to analyze signature amino acids that support the defined levels of classification. It estimates ML trees using RAxML, followed by branch annotation of amino acid substitutions. ALVIEW v1.25 was used to visualize and check signature amino acids.<sup>30</sup>

## 3 | RESULTS AND DISCUSSION

For each RSV subgroup, phylogenetic trees by maximum likelihood (ML) and Bayesian inferences were obtained for the alignments of complete genomes, concatenated SH-G-F genes (including their intergenic regions), the individual coding regions of the SH, G, and F genes, and the G-ectodomain and G 2nd hypervariable region (HR). However, convergence was not achieved for Bayesian tree of the RSV-B SH gene even after 50 million generations. The tree topologies of both ML and Bayesian inferences were similar (ML and Bayesian trees in Figures S1 and S2, respectively). Nevertheless, the longer the length of the region analyzed, the better the resolution of the trees, as seen when full genome sequences are used for the analyses. However, there is uneven representation of full-length global circulating strains both in number and in countries of origin, when compared to the availability of shorter fragments such as the G-ectodomain sequences (Figure S3).

The phylogenetic analysis of genomic regions reveals that most of them have high statistical support in the phylogenies' nodes ( $\geq 80\%$  bootstrap and  $\geq 0.8$  posterior probability values) with exception of

the trees obtained with the SH gene and the G 2nd HR. In addition, the plots of the absolute number of transitions and transversions vs divergence reveal loss of phylogenetic signal for both SH gene and the G 2nd HR, while most of the other analyzed regions show a constant increase in the absolute number of transitions and transversions with the increment of divergence (Figure S4).

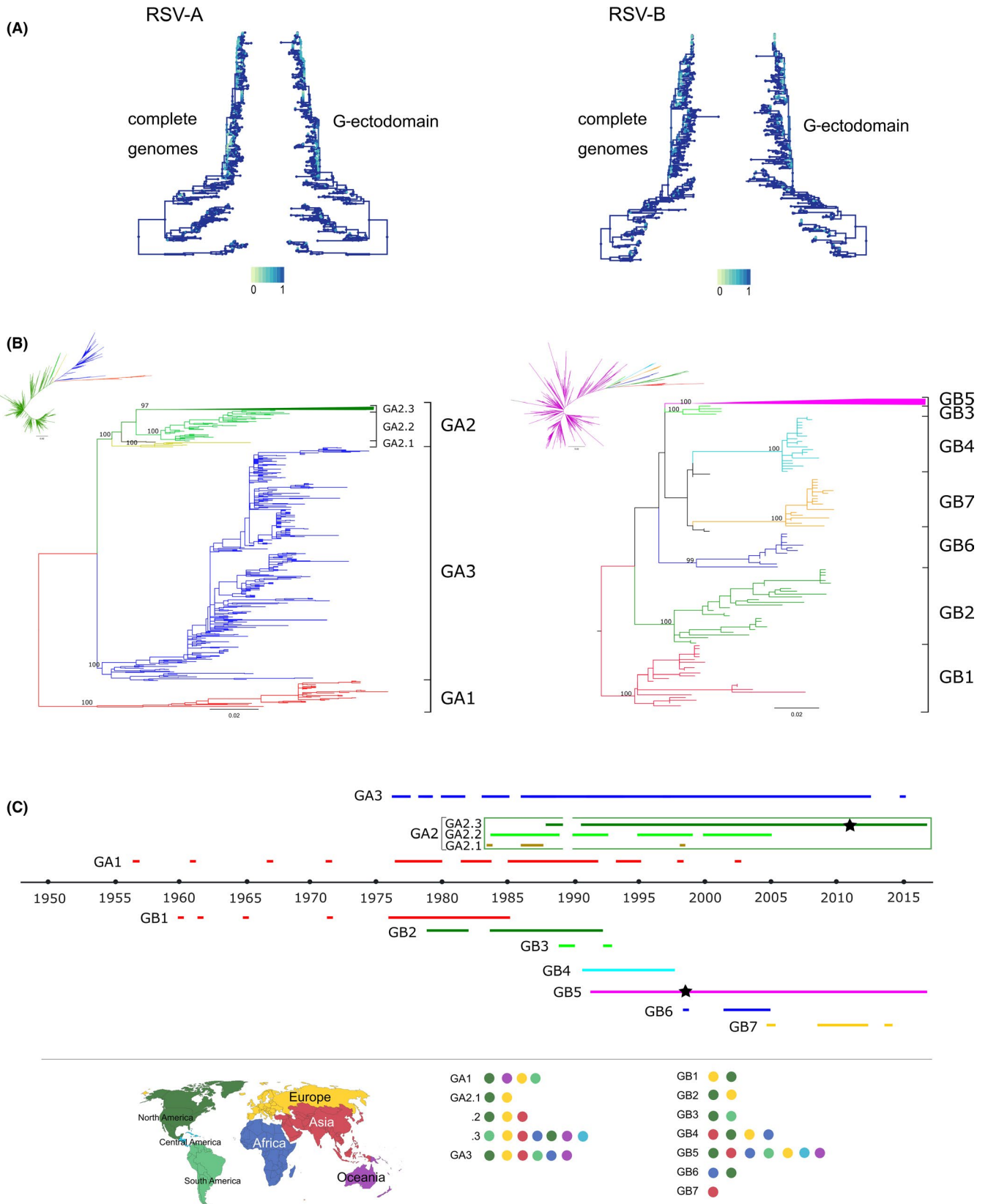
Phylogenies constructed from F gene, G gene, and G-ectodomain showed very similar topologies, albeit with lower resolution than complete genome trees. Given that vaccine and prophylaxis strategies are mostly targeted to the F protein, analysis of F sequences would provide useful data; however, there is more availability both in number and geographical representation of G-ectodomain sequences. Therefore, we focused on G-ectodomain. A comparison of the tree topology of the complete genome and the G-ectodomain phylogenies is shown in Figure 2A. Even though several clusters are not identical between complete genomes and G-ectodomain phylogenies, the detailed analysis by taxa shows that the inconsistencies occur largely in terminal nodes with low statistical support. The ancestor nodes of the main clades of sequences remain unchanged in their topology and statistical support, suggesting that potential genotype assignment of the sequences would not be influenced.

Based on these results, we conclude that the G-ectodomain should be considered the minimum region of the genome to be analyzed to obtain reliable phylogenies and genotype designation, and it can be used to obtain robust phylogeographic analyses allowing inferences about the origin of viral introductions in a particular geographical region.

Next, to classify genotypes, we based on the strategy of using average p-distances for RSV-A subgroup, described in 2015.<sup>17</sup> Briefly, in that work the average p-distances among individual genotypes, as well as within each genotype, were calculated. The highest intragenotype average p-distance was found for the GA1 genotype. This value was taken as the threshold for sorting viruses into different genotypes. Using these criteria, we reevaluated and redefined all previous genotypes as explained below and also summarized in Table S1.

One alignment of each RSV subgroup was obtained from all available curated G-ectodomain sequences up to February 2018. Phylogenetic analyses by two inference methods (ML and Bayesian) were performed. Monophyletic clades with high statistical support ( $\geq 80\%$  for bootstrap and  $\geq 0.8$  for posterior probabilities values) that clustered the oldest or first detected strains were identified and designated as GA1 for RSV-A and GB1 for RSV-B. A total of 44 non-identical G-ectodomain sequences (64 total sequences) were associated with GA1 and 22 non-identical G-ectodomain sequences (30 total

**FIGURE 2** Genotypes definition with G-ectodomain region. A, Comparison of tree backbones. Trees were constructed from an alignment of complete genome sequences per subgroup. Branches are colored according to each node's similarity to its best corresponding node in the opposite tree. In the scale, 0 means non-similarity and 1 means exact matching. B, Maximum likelihood trees for RSV-A and B with G-ectodomain sequence alignments (see the Section Methods for details). Colors in branches represent genotypes and subgenotypes. Subgenotype GA2.3 and genotype GB5 are collapsed in the rectangular trees. Complete uncollapsed trees are shown in the radial tree small figure next to each rectangular tree. UF bootstrap values of genotypes and subgenotypes nodes are shown. C, Timeline of detection of the identified genotypes and subgenotypes. Colored lines show the years of detection of a given genotype or subgenotype. Stars in lines of GA2 and GB5 genotypes show the year where strains with duplication in the 2nd hypervariable region of G gene were detected (72nt in RSV-A and 60nt in RSV-B). Geographical location of genotypes in order of detection is also shown



sequences) with GB1. The average intragenotype p-distance for GA1 was 0.036 (SE:0.004), and for GB1 was 0.032 (SE:0.004); thus, an intragenotype p-distance of 0.03 (3% divergence) was set as a general cut-off value for both subgroups. Based on this defined cut-off

value, the smallest number of well-supported genetic clades was identified and the average intraclade p-distance was calculated, and when the value resulted equal or below 0.03, the clade was defined as genotype (Tables 1A and 2). As a result, three genotypes were

identified for RSV-A (GA1-GA3) and seven genotypes for RSV-B (GB1-GB7). This collapses the previously larger number of groupings into a smaller set of genotypes for both RSV-A and RSV-B.

For a standardized genotype denomination, we propose to adopt the nomenclature defined by Peret et al (1998) where genotypes are designated as GAX and GBX; G stands for G-based genotype, A or B designate RSV subgroups, and X corresponds to a first-order ascending numbering system. This is a straightforward definition which does not allude to geographic references, avoiding potential stigmatizing labeling of RSV clades.<sup>6</sup> We renamed all genotypes in ascending order according to their first detection date, except GA2. We maintained its designation despite having emerged more recently than GA3, because it is a current widespread-circulating genotype and renaming this genotype could create confusion in the RSV community.

The minimum average intergenotype p-distance was 0.097 SE:0.008 (9% divergence) for RSV-A genotypes and 0.056 SE:0.006 (5% divergence) for RSV-B (Tables 1A and 2).

Small numbers of independent sequences not fitting the genotype definition were not classified and will remain undefined until future sequences cluster with them and meet the definition of

genotype. It is also possible that these sequences represent extinct viral genotypes.

Genotypes GA2, GA3, and GB5 are currently the most frequently detected and the largest ones (Figure 2B,C). Within these genotypes, further levels of classification were defined.

The next level of classification, defined as subgenotypes, encompassed well-supported dichotomies ( $\geq 80\%$  bootstrap and  $\geq 0.8$  posterior probability values) with an average intersubgenotype p-distance smaller than the minimum intergenotype p-distance described (divergence: 9% for RSV-A and 5% for RSV-B). The subgenotype nomenclature was defined as GAX.Y or GBX.Y, where X is the genotype and Y the subgenotype and corresponds to a second-order ascending numbering system. Within GA2, three subgenotypes were identified (GA2.1-GA2.3; Figure 2B) with average p-distances among them as shown in Table 1B. GA3 and GB5 did not show well-supported dichotomies within them. However, we cannot rule out the possibility that an increasing number of available sequences in the future will allow improved resolution of subgenotypes. For this reason, we propose to designate the level related to subgenotypes with a 0 (zero) until further resolution is possible.

**TABLE 1** Estimates of average genetic distances among and within RSV-A genotypes (A) and subgenotypes (B)

(A)			
	GA1 (44)	GA2 (2050)	GA3 (387)
GA1 (44)	<b>0.036 SE:0.004</b>		
GA2 (2050)	0.13 SE:0.01	<b>0.032 SE:0.003</b>	
GA3 (387)	0.13 SE:0.01	0.096 SE:0.009	<b>0.035 SE:0.003</b>
(B)			
	GA2.1 (8)	GA2.2 (52)	GA2.3 (1988)
GA2.1 (8)	<b>0.020 SE:0.003</b>		
GA2.2 (52)	0.058 SE:0.008	<b>0.028 SE:0.003</b>	
GA2.3 (1988)	0.074 SE:0.009	0.073 SE:0.009	<b>0.030 SE:0.003</b>

Note: Calculations were done for genotypes defined from a phylogenetic tree with 2481 unique RSV-A sequences of the G-ectodomain region. The number of sequences of each genotype (A) or subgenotype (B) is shown in parenthesis. P-distances were calculated between and among individual genotypes (A) or subgenotypes (B). P-distances within each genotype and subgenotype are denoted in bold. Standard error (SE) estimates are shown next to the average genetic distances.

**TABLE 2** Estimates of average genetic distances among and within RSV-B genotypes

	GB1 (22)	GB2 (28)	GB3 (4)	GB4 (21)	GB5 (1149)	GB6 (13)	GB7 (18)
GB1 (22)	<b>0.032 SE:0.004</b>						
GB2 (28)	0.072 SE:0.008	<b>0.037 SE:0.004</b>					
GB3 (4)	0.073 SE:0.008	0.064 SE:0.007	<b>0.015 SE:0.002</b>				
GB4 (21)	0.08 SE:0.01	0.077 SE:0.009	0.066 SE:0.009	<b>0.011 SE:0.002</b>			
GB5 (1149)	0.087 SE:0.008	0.077 SE:0.007	0.072 SE:0.008	0.083 SE:0.008	<b>0.032 SE:0.003</b>		
GB6 (13)	0.059 SE:0.007	0.056 SE:0.006	0.049 SE:0.007	0.059 SE:0.008	0.061 SE:0.008	<b>0.017 SE:0.004</b>	
GB7 (18)	0.074 SE:0.009	0.068 SE:0.008	0.064 SE:0.009	0.066 SE:0.009	0.075 SE:0.009	0.052 SE:0.007	<b>0.022 SE:0.003</b>

Note: Calculations were done for genotypes defined from a phylogenetic tree with 1259 unique RSV-B sequences of the G-ectodomain region. The number of sequences of each genotype is shown in parenthesis. P-distances within genotypes are denoted in bold. Standard error (SE) estimates are shown next to the average genetic distances.



Finally, in order to facilitate global surveillance, a further level of classification within subgenotypes was defined as lineages. To find an appropriate definition, we evaluated an alternative strategy to deal with the presence of polytomies in the trees, where lineages were defined according to the divergence of the strains from the ancestral node of the subgenotype. The allocation of lineages within each subgenotype was defined by the estimation of patristic distances (set as the sum of the lengths of the branches that link two nodes in a tree) between the internal nodes and the ancestral node of the subgenotype.

From the ML tree rooted with GA1 and GB1 genotypes, we visualized the nodes/taxa in increasing order of divergence. We assigned a patristic distance of 0 (zero) to the ancestral node of each subgenotype and set a cut-off of 0.015 patristic distance to define lineages (Figure 3). As we moved across the tree toward the terminal branches, when the node of a well-supported genetic clade showed a patristic distance  $\geq 0.015$  from the subgenotype ancestral node, it was defined as a new lineage. Subsequent lineages were therefore defined when diverged from the subgenotype ancestral node in multiples of 0.015 of patristic distance ( $>0.030$ ,  $>0.045$  and so on). The lineage nomenclature was defined as GAX.Y.Z or GBX.Y.Z where X is the genotype, Y the subgenotype, and Z corresponds to a third-order ascending numbering system.

As shown in Figure 3, six lineages were identified within GA2.3 subgenotype (GA2.3.1-GA2.3.6), five within GA3.0 (GA3.0.1-GA3.0.5), and five within GB5.0 (GB5.0.1-GB5.0.5).

When two or more divergent genetic clades, each having a distinct ancestral node, met the definition for new lineages within the same range of patristic distance, a lowercase letter after the lineage number was used to distinguish them. This was the case for GA3.0.3a and GA3.0.3b lineages, and GB5.0.4a, GB5.0.4b, and GB5.0.4c lineages. Once two or more particular lineages were defined with a letter, subsequent lineages were also defined with the same letter to denote the same ancestral origin, such as GB5.0.4a and GB5.0.5a (Figure 3C).

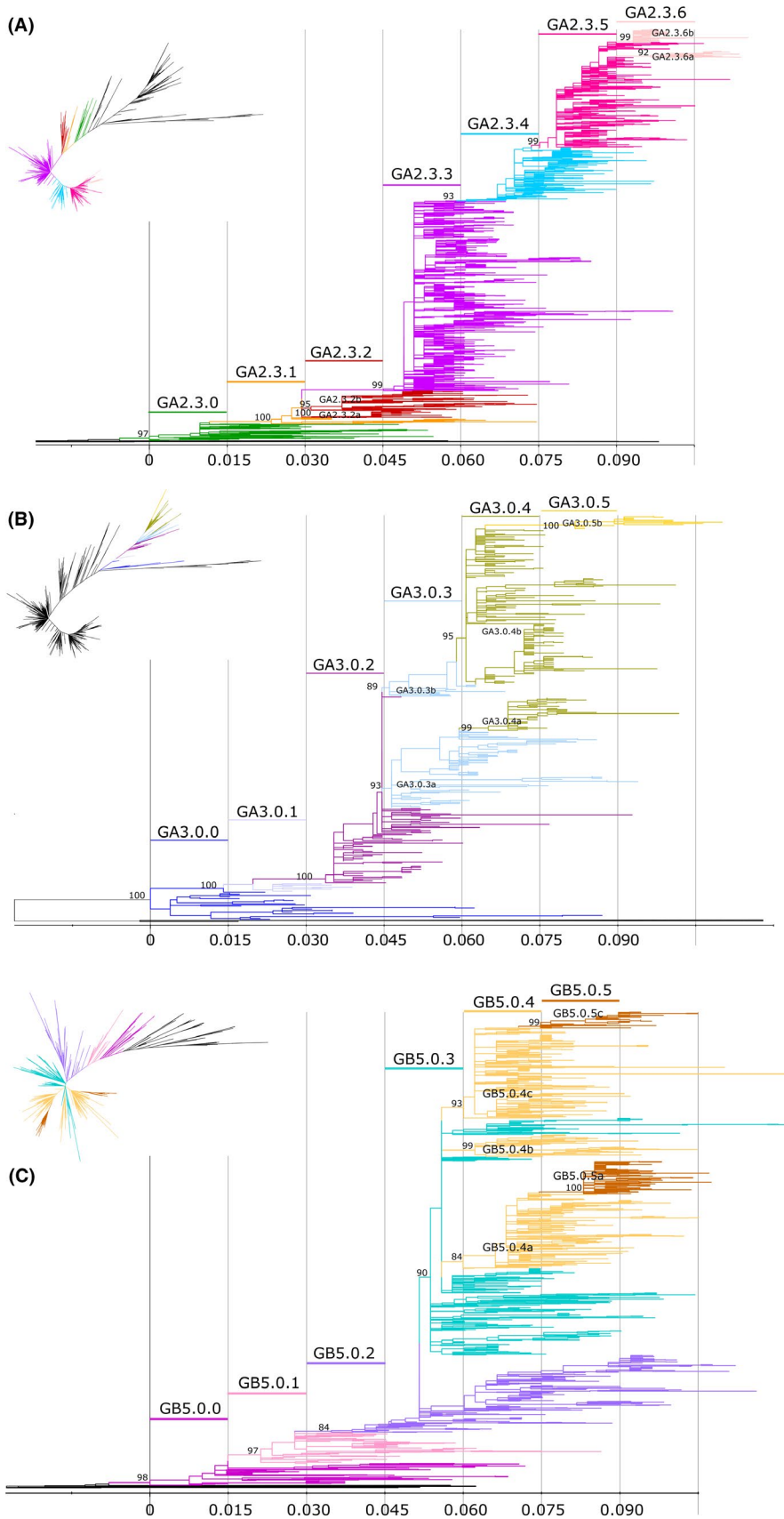
To complement the defined classification levels by phylogenetic clustering, we explored the search of signature amino acids, as used for influenza.<sup>31</sup> Signature amino acids are ideally defined by substitutions that are uniquely present in all sequences of a given clade. Table 3 and Figures S5 and S6 show amino acid substitutions across genotypes, subgenotypes, and lineages. Although most RSV-A and B genotypes, subgenotypes and lineages showed substitutions fitting this definition (defined as "main" in Table 3), there were other signature amino acids present in most but not all sequences within a clade, or present in more than one clade across the trees (defined as "secondary" in Table 3). Our analysis shows that all genotypes and subgenotypes in RSV-A and B were clearly defined by a collection of characteristic amino acid substitutions. For some lineages, there were no unique signature substitutions or at least two concomitant substitutions need to be present for a sequence to be properly classified into that lineage. Thus, we recommend that signature amino acids should be analyzed as a collection of substitutions (or haplotypes) defining each genotype/subgenotype/lineage and used as a

combined tool together with phylogenetic methods to further confirm the classification of sequences into given genetic clades.

With this new classification, the number of genotypes was reduced from 13 to three genotypes for RSV-A (GA1-GA3) and from 20 to seven genotypes for RSV-B (GB1-GB7). Two further levels of classification were added: subgenotypes and lineages (summarized definitions in Table S2). The timeline of detection of genotypes and subgenotypes is shown in Figure 2C. In addition, the information about period and regions of circulation, and the first detected strain for all the proposed taxonomic groups is listed in Table 4. For RSV-A, GA1 is the oldest genotype and was no longer detected after 2003. GA2 and GA3 genotypes currently circulate worldwide, but GA2 is the main circulating genotype and includes the strains with 72nt duplication in the 2nd HR of G (formerly named as ON1 strains) first described in 2011.<sup>11</sup> In the case of RSV-B, GB1 was first detected in 1960 and ceased its detection in 1985. GB5 and GB7 are the currently circulating genotypes, whereas GB6 was last detected in 2005. Only sequences from Thailand, Taiwan, China, and India clustered in GB7, showing geographical restriction within the Asian continent, unlike GB5, which shows a global circulation. Most GB5 sequences have the 60nt duplication in the 2nd HR of G, first described in 1999.<sup>15</sup> The schematic diagram of genotypes and subgenotypes circulation (Figure 2C) and the information listed in Table 4 are a snapshot and may not represent the real global circulation pattern of RSV due to bias in GenBank deposition practices.

For epidemiological purposes, the proposed procedure to classify newly sequenced strains into existing genotypes/subgenotypes/lineages is listed in Table S3. To facilitate the classification of any new sequence, curated reference alignments for both subgroups A and B are offered as supplementary materials (RSV-A-GectodomainReferenceAlignment and RSV-B-GectodomainReferenceAlignment). Figures S5 and S6 show ML trees obtained with these alignments (Bayesian inferences upon request). Both trees show congruence with ML trees inferred using the larger datasets for A and B depicted in Figures 2B and 3; therefore, they can be used for rapid classification of new sequences.

The instructions on how to define a new genotype/subgenotype or a lineage are summarized in Table S4 and are based on the procedures described above. As a caveat, the definition of new genotypes, subgenotypes, or lineages should be performed by using all the available G-ectodomain nucleotide sequences, because when sub-sampling datasets are obtained, the diversity of the lineages could be underrepresented. We propose that a global consortium of experts in RSV evolution working together with surveillance reference laboratories should be set up to assess the emergence of new genotypes, subgenotypes or lineages, or reassess existing nomenclature or definition on a regular basis. This consortium could form part of the International RSV Society which is a special interest group of the International Society for Influenza and other Respiratory Virus Diseases (isirv; <https://www.isirv.org/site/index.php/special-interest-groups/international-respiratory-syncytial-virus-society>). This consortium would be an ideal group to maintain updated reference



**FIGURE 3** Lineages identified in the subgenotype GA2.3, GA3.0 and GB5.0. Maximum likelihood trees in rectangular format showing lineages in subgenotypes: GA2.3 (A), GA3.0 (B) and GB5.0 (C). Colors in branches denote different lineages. The name of each lineage is shown above the colored line according to that lineage. Scale bar shows patristic distance by ranges of 0.015. UF Bootstrap of each ancestral lineage node is detailed. The small figure of the radial tree shows colored lineages in the uncollapsed whole trees for RSV-A or B, only the subgenotype lineages are denoted, the rest of the tree is colored in black



**TABLE 3** Signature amino acids characterizing genotypes/subgenotypes/lineages for RSV-A (A) and B (B)

(A)	
RSV-A genotypes	Signature amino acids
GA1	<b>101P,104P,111I,121G,123K,141A,157S,244I/T,258L,262M,280S,293P</b>
GA2	<b>I111P/S/F,P226L/H</b>
GA2.1	<b>K229N/S</b>
GA2.2	<b>V122A,[I114T + L115P+S117P/L]</b>
GA2.3	<b>V122A,P156Q,P289S/Y/F,S269T</b>
GA2.3.1	as GA2.3 plus <b>T252I,G254E</b>
GA2.3.2	as GA2.3 plus <b>T133I</b>
GA2.3.2a	as GA2.3.2 plus <b>N242S,V225I</b>
GA2.3.2b	as GA2.3.2 plus <b>T112I,I141M,R244K/N</b>
GA2.3.3	as GA2.3 plus <b>P111S,P292S/L</b>
GA2.3.4	as GA2.3 plus <b>F208L,N273Y/H</b>
GA2.3.5	as GA2.3 plus <i>insertion 261-284</i> , <b>E232G/R,T253K/R</b>
GA2.3.6a	as GA2.3.5 plus [ <b>V225A + L274P<sup>b</sup></b> ]
GA2.3.6b	as GA2.3.5 plus [ <b>P206Q + L274P/S<sup>b</sup></b> ]
GA3	<b>I111T,S250N</b>
GA3.0.0	<b>I111T,S250N</b>
GA3.0.1	<b>S102F,N191S,T295I</b>
GA3.0.2	as above plus <b>T113A,T125I,P274T</b>
GA3.0.3	as above plus <b>N161D,N297D</b>
GA3.0.3a	as GA3.0.3 plus <b>E232D</b>
GA3.0.4a	as GA3.0.3a plus <b>T227S</b>
GA3.0.3b	as GA3.0.3
GA3.0.4b	as GA3.0.3b plus <b>S217P,S224P,P246L,P206Q</b>
GA3.0.5b	as GA3.0.4b plus <b>N153K</b> .
(B)	
RSV-B genotypes	Signature amino acids
GB1	<b>137I,150S,152S,208I,222M,271S,280I</b>
GB2	<b>I118T,P223T,S257L</b>
GB3	<b>R136I</b>
GB4	<b>P216S,P219S,M222A,K224R,N230D,K234E,Q248R,K258D</b>
GB5	<b>T143N,L237P/S</b>
GB5.0.0	<b>T143N,L237P/S</b>
GB5.0.1	as above, plus <i>insertion 245-264</i> , <b>S247P<sup>b</sup></b>
GB5.0.2	as above, plus <b>T138S/F,K158-P159del<sup>a</sup>,L223P</b>
GB5.0.3	as above, plus <b>V271A</b>
GB5.0.4a	as GB5.0.3 plus [ <b>I200T + I281T</b> ]
GB5.0.5a	as GB5.0.4a plus <b>R136T,T254I<sup>b</sup></b>
GB5.0.4b	as GB5.0.3 plus [ <b>S267L + S297F/L</b> ] and <b>K153R</b> or <b>I111T</b>
GB5.0.4c	as GB5.0.3 plus <b>S291L</b> or <b>E305D</b>
GB5.0.5c	as GB5.0.4c subclade <b>S291L</b> , plus <b>R262G<sup>b</sup></b>

(Continues)

**TABLE 3** (Continued)

(B)	
RSV-B genotypes	Signature amino acids
GB6	<b>A116T,T126K,P235L,Q262L/P</b>
GB7	<b>Q180R,R214I,T239I,R242K,V251M/I,E272D</b>

Note: Amino acid annotation follows the format "X122Y," where X = reference, Y = signature, number = position from beginning of G-ORF. Signature amino acids are defined against the consensus sequence of GA1 for RSV-A and GB1 for RSV-B. "Main" amino acid signatures are in bold, "secondary" changes are in italics. Substitutions in square brackets are needed in combination to be considered as signature amino acids.

<sup>a</sup>alternative position for this deletion: P159-K160del.

\*Substitutions located within the duplication region of the G gene (72nt in RSV-A or 60nt in RSV-B). For these substitutions, the positions were described using strains JN257693\_CAN\_2010 (RSV-A) and DQ227364\_ARG\_1999 (RSV-B) as reference sequences.

datasets using curated reference sequences representing each of the defined classification levels, to be used by the global community.

### 3.1 | Strengths and weaknesses of this RSV strain classification proposal

All of the analyses in the present study were carried out with sequences downloaded from GenBank, which is not a dedicated and curated database. A variety of issues with the sequences were found that might have affected tree topologies and calculation of average p-distances. Therefore, a reliable well-maintained database with curated sequences will be essential to standardize RSV molecular surveillance.

Our study shows that full genome sequences are the most informative and desirable dataset for genotyping purposes. As technology progresses and NGS methodologies become widespread in the near future, costs will eventually reduce enabling more laboratories to implement protocols to generate full viral genomes for their molecular analyses. The strength of this proposal is that genotype, subgenotype, and lineage definitions proposed here are scalable to the complete genome. The cut-off value for the average p-distance should be recalculated as shown in Table S5, and patristic distances should be evaluated for the identification of lineages. The strategy proposed in this work relies on the current availability of partial RSV sequences. The results of this study and other publications support the idea that the G-ectodomain sequences can be effectively used for genetic characterization of RSV strains.<sup>32-34</sup>

It is important to highlight the economic benefit of a low-cost methodology especially for low- and middle-income countries that may be expanding their sequencing activity, given that the length of G-ectodomain region can be sequenced by Sanger methodology with only two reactions, allowing surveillance laboratories to monitor molecular diversity easily and in real time. In this effort to standardize the molecular classification of RSV strains, we consider that the use of a cut-off value for genotype definition is essential.

**TABLE 4** Information about period and regions of circulation, and the first detected strain for all the proposed taxonomic groups

Subgroup	Genotype	Subgenotype	Lineage	Period detected <sup>a</sup>	Regions of circulation <sup>b</sup>	First detected strain	
A	1			1956-2003	North America, South America, Europe, Oceania	AY911262	
		2	2.1	1984-1998	North America, Europe	KP258733	
			2.2	1984-2005	Asia, North America, Europe	HQ731720	
		2.3	2.3.0	1977-2012	Asia, Africa, North America, South America, Europe	KU316166	
			2.3.1	1997-2009	Asia, Africa, North America, South America, Europe, Oceania	AF193324	
			2.3.2a	1999-2006	Africa, South America, Europe	AY343612	
			2.3.2b	1999-2008	Asia, Africa, North America, South America, Europe	AY343603	
			2.3.3	2004-2015	Global	MF496446	
			2.3.4	2003-2016	Global	JX513373	
			2.3.5	2010-2016	Global	JN257693	
			2.3.6a	2011-2015	South America, Europe	JX513365	
			2.3.6b	2013-2015	Africa	KX453341	
		3	3.0	3.0.0	1976-1997	Asia, North America, Europe	HQ731715
				3.0.1	1982-1990	North America, Europe	MG642031
				3.0.2	1989-2003	Asia, North America, South America, Europe	AY343587
				3.0.3a	1998-2011	Asia, North America, South America, Europe, Oceania	JX069802
				3.0.3b	2000-2006	North America, South America, Europe	AY343561
				3.0.4a	1999-2009	South America, Europe	EU025205
				3.0.4b	2000-2013	Asia, Africa, North America, South America, Europe, Oceania	AY343558
				3.0.5b	2006-2015	Asia, Africa, North America, South America, Europe, Oceania	JX645865
B	1			1960-1985	North America, Europe	M73545	
	2			1979-1993	North America, Europe	KP258712	
	3			1989-1993	North America, South America	M73543	
	4			1991-1997	Asia, Africa, North America, Europe	AF193332	
	5	5.0	5.0.0	1992-2012	Asia, Africa, North America, South America, Europe	KP258745	
			5.0.1	1999-2012	Asia, Africa, North America, South America, Europe	AY333364	
			5.0.2	2004-2015	Global	JX489436	
			5.0.3	2005-2014	Global	JX576756	
			5.0.4a	2008-2015	Asia, Africa, North America, South America, Europe, Oceania	MF496630	
			5.0.4b	2004-2011	Asia, Africa, America, Europe	DQ227395	
			5.0.4c	2007-2015	Global	KC297480	
			5.0.5a	2013-2016	Asia, Africa, North America, South America, Europe, Oceania	KY249660	
		5.0.5c	2010-2016	Asia, Africa	KF246600		
	6			1998-2005	Africa, North America	JF704213	
7			2005-2014	Asia	MF496621		

<sup>a</sup>Period detected up to February 2018.<sup>b</sup>This may not represent the current real global circulation pattern of RSV due to bias in GenBank deposition practices.

Defining the average intragenotype p-distance of GA1 and GB1 as cut-off is very useful because these clades include the oldest strains detected for each subgroup and they also are no longer detected in the population. Although unlikely, we cannot rule out the possibility that in the near future, strains from these two genotypes might still be detected. Genetic distance measure is sensitive to diversity of sampled sequences; thus, if more sequences from old strains were to be released in the future, this could alter average p-distances of intra/interclades and even slightly modify the cut-off value. We strongly reinforce the need for periodical reevaluation of genotype definitions by a global consortium of RSV experts.

## 4 | CONCLUSION

With many new vaccine candidates in prospect, widespread adoption of a unified criterion for RSV genotyping will enable standardized, comparable analyses across the scientific research and public health communities. Thus, our proposal of RSV strain classification will facilitate the analysis of strains in clinical and epidemiological studies and would be a fitting start to a joint global RSV surveillance.

## ACKNOWLEDGEMENTS

We would like to thank Burcu Ermetal for assistance with the use of the TREESUB program and interpretation of data.

## CONFLICT OF INTEREST

The authors declare that they have no competing interests.

## DEDICATION

We would like to dedicate this manuscript to the memory of Dr José Antonio Melero, who was an outstanding researcher and warm and generous individual who encouraged us to work on this topic, with whom we began to work on the unification of the RSV genotype definition shortly before his untimely death.

## ORCID

Stephanie Goya  <https://orcid.org/0000-0001-7479-3064>

Mariana Viegas  <https://orcid.org/0000-0002-6506-1635>

## REFERENCES

- Shi T, McAllister DA, O'Brien KL. Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015: a systematic review and modelling study. *Lancet Lond Engl*. 2017;390:946-958.
- Neuzil KM. Progress toward a respiratory syncytial virus vaccine. *Clin Vaccine Immunol*. 2016;23(3):186-188.
- Rima B, Collins P, Easton A, et al. ICTV virus taxonomy profile: pneumoviridae. *J Gen Virol*. 2017. 98(12):2912-2913.
- Zlateva KT, Lemey P, Moës E, Vandamme AM, Van Ranst M. Genetic variability and molecular evolution of the human respiratory syncytial virus subgroup B attachment G protein. *J Virol*. 2005. 79(14):9157-9167.
- Collins PL, Karron RA. *Respiratory Syncytial Virus and Metapneumovirus Fields Virology 6th ed*. Philadelphia, PA: Lippincott Williams & Wilkins; 2013.
- Peret TC, Hall CB, Schnabel KC, Golub JA, Anderson LJ. Circulation patterns of genetically distinct group A and B strains of human respiratory syncytial virus in a community. *J Gen Virol*. 1998;79(Pt 9):2221-2229.
- Peret TCT, Hall CB, Hammond GW, et al. Circulation patterns of group A and B human respiratory syncytial virus genotypes in 5 communities in North America. *J Infect Dis*. 2000;181:1891-1896.
- Venter M, Madhi SA, Tiemessen CT, Schoub BD. 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*. 2001;82:2117-2124.
- Cui G, Zhu R, Qian Y, 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*. 2013;8(9):e75020.
- Shobugawa Y, Saito R, Sano Y, et al. Emerging genotypes of human respiratory syncytial virus subgroup A among patients in Japan. *J Clin Microbiol*. 2009;47:2475-2482.
- Eshaghi AliReza, Duvvuri VR, Lai R, 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*. 2012;7(3):e32807.
- Hirano E, Kobayashi M, Tsukagoshi H, et al. Molecular evolution of human respiratory syncytial virus attachment glycoprotein (G) gene of new genotype ON1 and ancestor NA1. *Infect Genet Evol*. 2014;28:183-191.
- Blanc A, Delfraro A, Frabasile S, Arbiza J. Genotypes of respiratory syncytial virus group B identified in Uruguay. *Arch Virol*. 2005;150:603-609.
- Dapat IC, Shobugawa Y, Sano Y, et al. New genotypes within respiratory syncytial virus group B genotype BA in Niigata, Japan. *J Clin Microbiol*. 2010;48:3423-3427.
- Trento A, Viegas M, Galiano M, 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*. 2006;80:975-984.
- Trento A, Galiano M, Videla C, et al. Major changes in the G protein of human respiratory syncytial virus isolates introduced by a duplication of 60 nucleotides. *J Gen Virol*. 2003;84:3115-3120.
- Trento A, Ábrego L, Rodríguez-Fernández R, 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*. 2015;89:7776.
- Dimitrov KM, Abolnik C, Afonso CL, et al. Updated unified phylogenetic classification system and revised nomenclature for Newcastle disease virus. *Infect Genet Evol*. 2019;74:103917.
- Bankamp B, Byrd-Leotis LA, Lopareva EN, et al. Improving molecular tools for global surveillance of measles virus. *J Clin Virol*. 58(1):176-182.
- Smith DB, Bukh J, Kuiken C, et al. Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource. *Hepatology*. 59(1):318-327.
- Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *NucleicAcids Res*. 2004;32:1792-1797.
- Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015;32:268-274.
- Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermini LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017;14:587-589.

24. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol.* 2018;35:518-522.
25. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* 2018;4:vey016.
26. Robinson O, Dylus D, Dessimoz C. Phylo.io: interactive viewing and comparison of large phylogenetic trees on the web. *Mol Biol Evol.* 2016;33:2163-2166.
27. Rambaut A, Welch J, Dudas G, et al. FigTree. 2016.
28. Xia X. DAMBE7: New and improved tools for data analysis in molecular biology and evolution. *Mol Biol Evol.* 2018;35:1550-1552.
29. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol.* 2016;33:1870-1874.
30. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics.* 2014;30:3276-3278.
31. European Centre for Disease Prevention and Control. Influenza virus characterisation, February 2018 30.
32. Rojo GL, Goya S, Orellana M, et al. Unravelling respiratory syncytial virus outbreaks in Buenos Aires, Argentina: Molecular basis of the spatio-temporal transmission. *Virology.* 2017;508:118-126.
33. Viegas M, Goya S, Mistchenko AS. Sixteen years of evolution of human respiratory syncytial virus subgroup A in Buenos Aires, Argentina: GA2 the prevalent genotype through the years. *Infect Genet Evol.* 2016;43:213-221.
34. Otieno JR, Kamau EM, Oketch JW, et al. Whole genome analysis of local Kenyan and global sequences unravels the epidemiological and molecular evolutionary dynamics of RSV genotype ON1 strains. *Virus Evol.* 2018;4(2):vey027.

#### SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Goya S, Galiano M, Nauwelaers I, et al. Toward unified molecular surveillance of RSV: A proposal for genotype definition. *Influenza Other Respi Viruses.* 2020;00:1-12. <https://doi.org/10.1111/irv.12715>