Standardized Phylogenetic Classification of Human Respiratory Syncytial Virus below the Subgroup Level

Stephanie Goya,[1] Christopher Ruis,[2] Richard A. Neher,[2] Adam Meijer, Ammar Aziz, Angie S. Hinrichs, Anne von Gottberg, Cornelius Roemer, Daniel G. Amoako,[3] Dolores Acuña, Jakob McBroome, James R. Otieno, Jinal N. Bhiman, Josie Everatt, Juan C. Muñoz-Escalante, Kaat Ramaekers,[4] Kate Duggan, Lance D. Presser, Laura Urbanska, Marietjie Venter, Nicole Wolter, Teresa C.T. Peret, Vahid Salimi, Varsha Potdar, Vítor Borges, Mariana Viegas[1]

A globally implemented unified phylogenetic classification for human respiratory syncytial virus (HRSV) below the subgroup level remains elusive. We formulated global consensus of HRSV classification on the basis of the challenges and limitations of our previous proposals and the future of genomic surveillance. From a high-quality curated dataset of 1,480 HRSV-A and 1,385 HRSV-B genomes submitted to GenBank and GISAID (https://www.gisaid.org) public sequence databases through March 2023, we categorized HRSV-A/B sequences into lineages based on phylogenetic clades and amino acid markers. We defined 24 lineages within HRSV-A and 16 within HRSV-B and provided guidelines for defining prospective lineages. Our classification demonstrated robustness in its applicability to both complete and partial genomes. We envision that this unified HRSV classification proposal will strengthen HRSV molecular epidemiology on a global scale.

Human respiratory syncytial virus (HRSV) is a leading cause of acute lower respiratory tract infection in children, elderly, and immunocompromised persons. In 2023, the US Food and Drug Administration and the European Medicines Agency approved the first HRSV vaccines (*1*,*2*). Simultaneously, a monoclonal antibody was approved for widespread use in infants and not limited to high-risk and premature children (*3*). The availability of HRSV immunization highlights the role of molecular epidemiology

Author affiliations: University of Washington, Seattle, Washington, USA (S. Goya); University of Cambridge, Cambridge, UK (C. Ruis); University of Basel and SIB, Basel, Switzerland (R.A. Neher, C. Roemer, L. Urbanska); National Institute for Public Health and the Environment, Bilthoven, the Netherlands (A. Meijer, L.D. Presser); World Health Organization Collaborating Centre for Reference and Research on Influenza, Melbourne, Victoria, Australia (A. Aziz); University of California Santa Cruz, Santa Cruz, California, USA (A.S. Hinrichs, J. McBroome); National Institute for Communicable Diseases of the National Health Laboratory Service, Johannesburg, South Africa (A. von Gottberg, J.N. Bhiman, J. Everatt, N. Wolter); University of Witwatersrand, Johannesburg, South Africa (A. von Gottberg, J.N. Bhiman, N. Wolter); University of KwaZulu-Natal, Durban, South Africa (D.G. Amoako); Universidad Nacional de La Plata, Buenos Aires, Argentina (D. Acuña, M. Viegas); National Scientific and Technical Research Council, Buenos Aires, Argentina (D. Acuña, M. Viegas); Theiagen Genomics, Highlands Ranch, Colorado, USA (J.R. Otieno); Autonomous University of San Luis Potosí, San Luis Potosí, Mexico (J.C. Muñoz-Escalante); Rega Institute for Medical Research, Leuven, Belgium (K. Ramaekers); University of Edinburgh, Edinburgh, Scotland, UK (K. Duggan); University of Pretoria, Pretoria, South Africa (M. Venter); University of Texas Medical Branch, Galveston, Texas, USA (T.C.T. Peret); Tehran University of Medical Sciences, Tehran, Iran (V. Salimi); ICMR National Institute of Virology, Pune, India (V. Potdar); National Institute of Health Doutor Ricardo Jorge, Lisbon, Portugal (V. Borges)

DOI: http://doi.org/10.3201/eid3008.240209

[1]These authors were co–principal investigators.

[2]These authors contributed equally to this article.

[3]Current affiliation: Department of Pathobiology, University of Guelph, Guelph, Ontario, Canada.

[4]Current affiliation: Sciensano, Infectious Diseases in Humans, Unit (Re)-Emerging Viruses, Brussels, Belgium.

as a tool to monitor their efficacy. Standards for HRSV nomenclature for sharing of viral isolates and sequences in databases have been published (*4*). Nevertheless, a standardized HRSV phylogenetic classification system has yet to be defined and implemented.

In 2022, HRSV was designated as *Orthopneumovirus hominis* species within the Pneumoviridae family. Below species level are 2 antigenic groups, known as HRSV subgroup A (HRSV-A) and B (HRSV-B), that were previously referred to as subtypes (*4*–*6*). Within each subgroup, genotypes were initially defined based on statistically supported phylogenetic clades inferred with the second hypervariable region of the G gene (Figure 1, panels A, B) (*7*). The G gene, encoding the attachment glycoprotein, exhibits the highest genetic and antigenic variability. Of note, the gene has undergone a duplication of a 72-nt fragment in HRSV-A and 60-nt fragment in HRSV-B (Figure 1, panel B) (*8*,*9*).

To identify emerging genotypes, researchers have used genetic distances between phylogenetic clades and distinctive genetic features, accompanied by variable nomenclature based on the gene (GA1–GA7 in HRSV-A and GB1–GB4 in HRSV-B), country and subgroup (SAB1–SAB4 for South African genotypes in HRSV-B), or city and province (NA1–NA2 [Niigata] and ON1 [Ontario] in HRSV-A, BA1–BA9 [Buenos Aires] in HRSV-B) (*7*–*16*). Since 2020, alternative phylogenetic reclassifications have been proposed; Goya et al. established a hierarchical classification system for HRSV phylogenies, comprising genotypes, subgenotypes, and lineages, using the G gene (*17*). That framework enabled laboratories without capacity for whole-genome sequencing to conduct molecular epidemiology studies. Independently, Ramaekers et al. (*18*) proposed reclassifications into lineages and Chen et al. (*19*) into genotypes using complete HRSV genomes. Those approaches support comprehensive monitoring of viral evolution across all genes, including the F gene encoding the fusion protein, a crucial target for monoclonal antibodies and the foundation of approved HRSV vaccines (Figure 1, panel A). Of note, challenges in HRSV molecular epidemiology persisted within the reclassification-defined categories because of reliance on genetic or patristic distances between tree tips or nodes.

The milestones achieved in HRSV interventions have renewed interest in addressing the challenge of

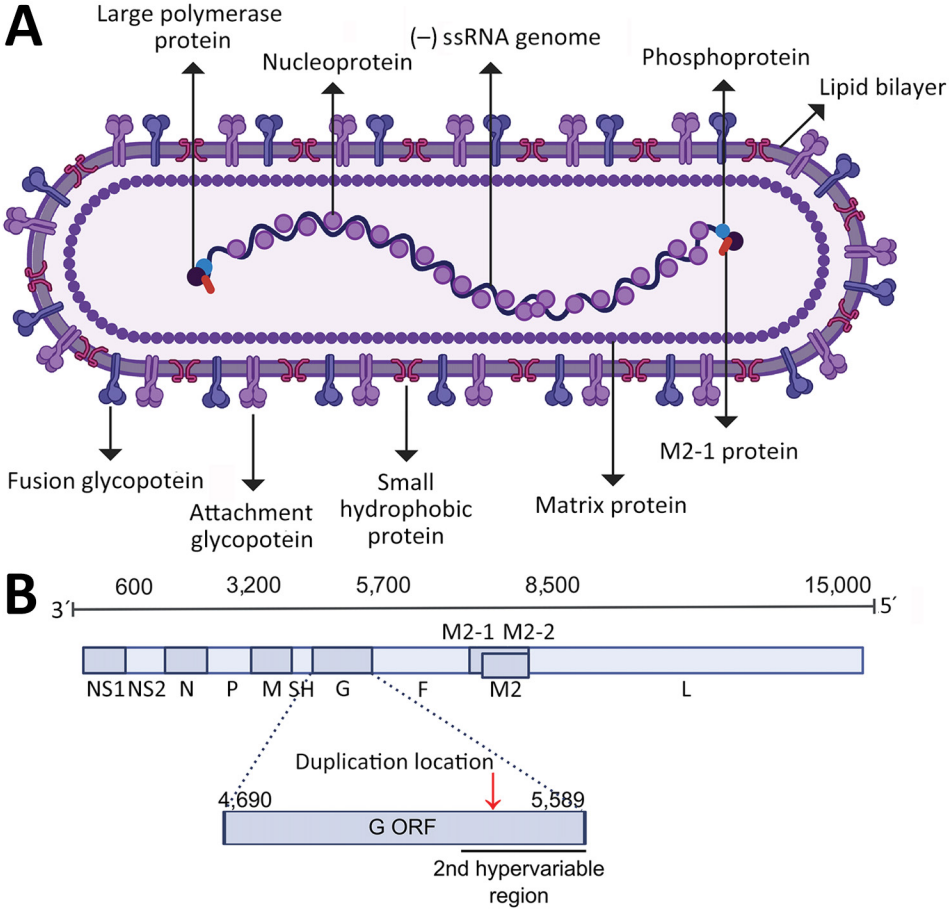


Figure 1. The structure and genome of human respiratory syncytial virus (HRSV). A) Schematic of the HRSV virion structure detailing the location of structural proteins. B) Schematic of the HRSV genome organization with the approximated location of genes highlighted; the exact location slightly differs between subgroups and strains. The location of the second hypervariable region in the G gene, used originally for molecular epidemiology classification, is detailed. Red arrow in panel B indicates location of the G gene 72-nt duplication in HRSV-A and 60-nt duplication in HRSV-B. Figure created with BioRender (https://www.biorender.com). F, fusion glycoprotein; G, attachment glycoprotein; L, large polymerase protein; M, matrix protein; M2, M2 protein; N, nucleocapsid; NS, nonstructural protein; ORF, open reading frame; P, phosphoprotein; SH, small hydrophobic protein.

classifying HRSV below the subgroup level. Those advances prompted establishment of the HRSV Genotyping Consensus Consortium (RGCC), formed by HRSV and virus evolution experts aiming to provide standardized criteria for harmonizing global HRSV molecular surveillance efforts. We present a novel framework for HRSV classification below the subgroup level, based on current knowledge of HRSV diversity and evolution, focused on practical implementation for molecular epidemiology.

Methods

HRSV Sequences Dataset

We downloaded HRSV complete genomes from the National Center for Biotechnology Information Virus (https://www.ncbi.nlm.nih.gov/labs/virus) and GISAID EpiRSV (https://www.gisaid.org) databases through March 11, 2023, using a filter for sequence length >14,000 nt, obtained from human hosts and including the year and country of the sample collection (Appendix 1 Figure 1, https://wwwnc.cdc.gov/EID/article/30/8/24-0209-App1.pdf). We reserved sequences containing nucleotide ambiguities, indicating inadequate sequencing depth, for epidemiologic analysis but excluded them from formal lineage definition (Appendix 1).

We aligned sequences with MAFFT version 7.490, and inspected and corrected alignment artifacts with Aliview version 1.28 (https://ormbunkar.se/aliview), mainly in the G gene (*20*,*21*). We trimmed alignment ends to encompass complete genomes from the first codon of the first gene (NS1) to the last codon of the last gene (L). We considered partial genomes if the lack of sequence was within 50 nt of the genome ends. We used RSVsurver (https://rsvsurver.bii.a-star.edu.sg) to identify and remove genomes with nucleotide insertions or deletions causing frameshift in any open reading frame. After alignment trimming, detection of identical sequences prompted redundancy removal using BBmap (https://jgi.doe.gov/data-and-tools/software-tools/bbtools), resulting in the final set of 1,538 HRSV-A and 1,387 HRSV-B genomes (Appendix 1 Figure 1).

Phylogenetic Analysis

We constructed maximum-likelihood phylogenetic trees with IQ-TREE version 2.2.0 (http://www.iqtree.org) (Appendix 1). We considered monophyletic clades statistically supported when SH-aLRT value was ≥80% and UFBoot2 value was ≥90% (*22*,*23*) (Appendix 1). We assessed temporal signal with TempEst version 1.5.3 (http://tree.bio.ed.ac.uk/software/tempest), and we inferred molecular-clock phylogenies with TreeTime (https://github.com/neherlab/treetime) (*24*).

We inferred the ancestral sequence reconstruction using Augur bioinformatic toolkit version 23.1.0 (https://docs.nextstrain.org/projects/augur/en/23.1.0) (*25*). We assessed recombination events by alignment-based method using RDP4 (http://web.cbio.uct.ac.za/~darren/rdp.html) and phylogenetic-based TreeKnit (https://pierrebarrat.github.io/TreeKnit.jl) (Appendix 1). We inferred the amino acid substitutions linked to the clades in the tree using Augur and automated the initial screening of lineages with Autolin (*26*). We manually curated amino acid comparison among monophyletic clusters to rectify conflicts arising from internal (nested) lineages and the confirmation of the lineage-defining amino acids in >90% of the clade's sequences. Results are available at https://github.com/rsv-lineages/Classification_proposal.

Results

Baseline Agreements on the HRSV Classification Definition

Our proposed classification establishes HRSV lineages for viruses below subgroup level. Studies have shown that HRSV phylogenetic trees constructed with complete genomes exhibit superior resolution (*17*–*19*). Therefore, we defined a classification system based on maximum-likelihood phylogenetic trees inferred from complete HRSV genomes. The maximum-likelihood algorithm formulates hypotheses about the evolutionary relationships among sequences; the implementation within IQ-TREE dealing with large datasets makes it particularly well suited to assert HRSV genomic phylogeny including sequences collected >50 years ago (*22*). We defined complete HRSV genomes to the nucleotide sequences spanning from the first codon of the first gene (NS1) to the last codon of the last gene (L). We considered almost-complete genomes if the sequence information gaps were within a 50-nt window at the genome ends. To define lineages, we only used genomes without nucleotide ambiguities (in accordance with the IUPAC code for nucleotide degeneracy).

Genomic Dataset Used for Lineages Definition

Applying the established baseline agreements, we gathered 1,538 HRSV-A and 1,387 HRSV-B high-quality genomes from public databases. The dataset revealed a limited global HRSV genomic surveillance; <20 genomes deposited annually through 2007 (Figure 2, panel A; Appendix 1 Figure 2). Since 2008, the

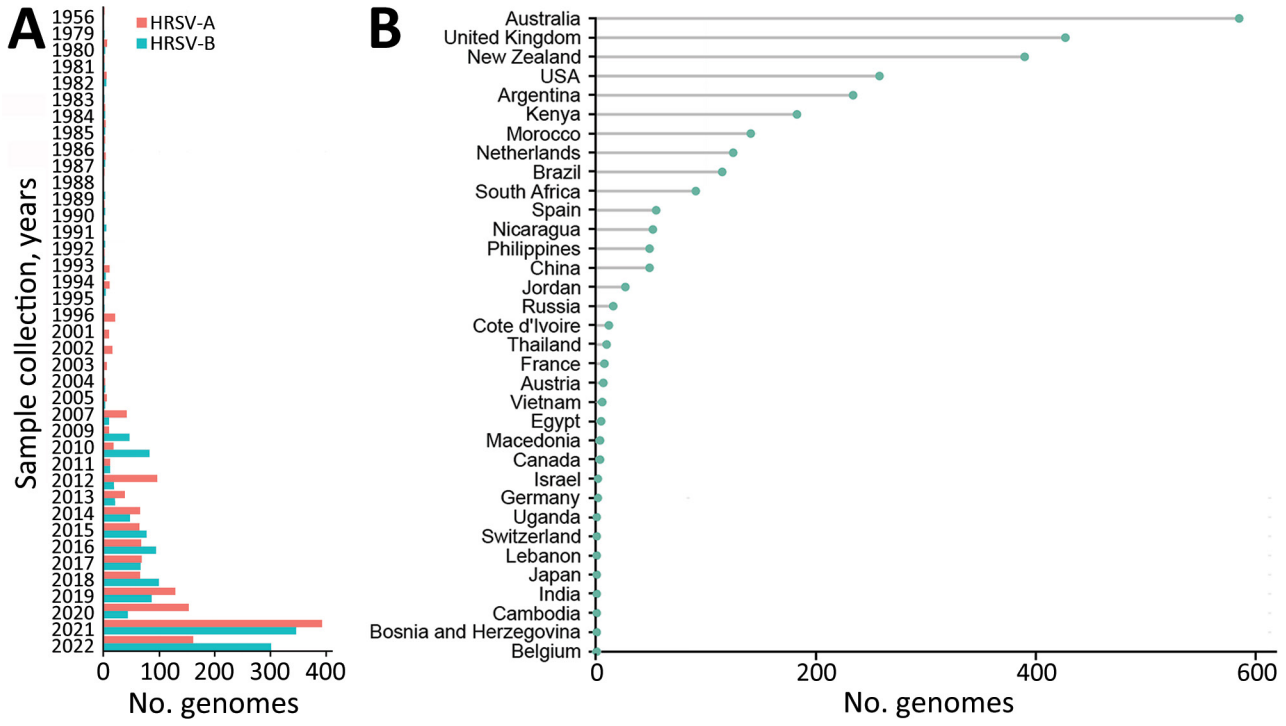


Figure 2. The global HRSV genomics surveillance landscape. HRSV genomes from GenBank and GISAID (https://www.gisaid.org) databases through March 11, 2023, that met inclusion criteria used for classification are shown by year of sample collection and subgroup (A) and by country of origin (B). HRSV, human respiratory syncytial virus.

number of genomes and representation of countries improved; a surge occurred after 2021, probably driven by expansion of viral genomics since the SARS-CoV-2 pandemic and the approval of the HRSV prophylactic treatments (Figure 2, panel A; Appendix 1 Figure 2). Considering delays in genome deposition in public databases, the number of genomes in 2022 may be higher than those used in this study. Regarding geographic representation, 9 countries (Australia, United Kingdom, New Zealand, United States, Argentina, Kenya, Morocco, Netherlands, and Brazil) submitted >100 genomes; only the United Kingdom achieved uninterrupted surveillance since 2008, but Australia deposited the most genomes globally (Figure 2, panel B).

Accurate Root Placement in HRSV Phylogenetic Trees

We reconstructed maximum-likelihood phylogenetic trees for the HRSV-A and HRSV-B datasets. We used 2 approaches to root the trees: the use of an outgroup, a conventional method for inferring the tree root using sequences known to be evolutionarily distant; and phylodynamic analysis, integrating temporal and phylogenetic patterns in virus evolution (Appendix 1). Both approaches consistently identified the same root for each subgroup cluster (Appendix 1 Figure 3). Phylodynamic analysis also identified 58 outlier sequences for HRSV-A and 2 for HRSV-B that were excluded from lineage designation. The final dataset considered for lineage designation comprised 1,480 HRSV-A and 1,385 HRSV-B genomes (Appendix 2 Table, https://wwwnc.cdc.gov/EID/article/30/8/24-0209-App2.xlsx).

HRSV Lineage Definition

We defined HRSV lineage as a statistically supported monophyletic cluster comprising ≥10 sequences and characterized by ≥5 aa substitutions, compared to the parental lineage. The lineage-defining amino acids, present in ≥90% of the sequences within the clade, may be found in any of the viral proteins.

Phylogenetic classifications vary among viral species aiming to define clusters reflecting the heterogeneity of the viral population, considering each virus unique evolutionary characteristics and using arbitrary thresholds for long-term applicability (*27*–*29*). Inherent bias exists in any classification system because of availability and spatiotemporal representation sequences. Therefore, our HRSV lineage definition did not include criteria of sequences from different outbreaks or countries to enable early detection of novel lineages. However, we propose establishing a threshold of ≥10

genomes for defining a lineage to monitor HRSV strains circulating within communities.

We observed the presence of distinctive signature amino acids shared by sequences of a phylogenetic clade in comparison to the parental lineage is a simple method to identify a new lineage. Methods (i.e., average nucleotide genetic distances, average patristic distances, or patristic distances between nodes) need phylogenies with complete datasets to define new categories, becoming complex with rapid increases of available sequences (*16*–*19*). In our proposal, we initially screened different amino acid thresholds in an automated manner, ranging from 1–10 lineage-defining amino acids (Appendix 1). The number of small lineages decreased as the number of lineage-defining amino acids increased, and 5 amino acids resulted in an intermediate complexity of lineages defined for both HRSV subgroups. Furthermore, we proposed that the lineage-defining amino acids should be conserved in ≥90% of the genomes within a clade, considering the potential reversion in some of the genomes within highly mutated hotspot sites. We acknowledged that other numbers of genomes or amino acids thresholds could be useful, but we emphasized that the key to establishing a global consensus is clear operational guidelines and a robust classification, 2 aspects that our proposal fulfills.

HRSV Lineage Nomenclature

We defined the lineage nomenclature integrating the HRSV subgroup letter and ascending ordinal numbers, separated by dots to represent nested lineages (Figure 3, panels A, B; Figure 4, panels A, B). Furthermore, we assigned a distinct nomenclature to the 72-nt (24-aa) G-gene duplication within HRSV-A and 60-nt (20-aa) G-gene duplication within HRSV-B. Those genetic events are epidemiologically relevant, because only viruses with G-gene duplication have been detected since 2017 (*30*–*33*). To track those viruses, we used the alias D, specifically A.D (historically, ON1 genotype) for HRSV-A and B.D (historically, BA genotype), for HRSV-B and nested lineages with increasing ordinal numbers. In summary, letters A and B indicate the HRSV subgroup at the beginning of the lineage name, C is unused, and D serves as an alias for 72-nt and 60-nt duplication within the G gene. In addition, aliases starting from E are limited to 3 numerical levels of nested lineages, preventing indefinite accumulation of numbers. For example, B.D.4.1.1 lineage has descendant lineages named B.D.E.1–B.D.E.4 instead of B.D.4.1.1.1–B.D.4.1.1.4, where E represents 4.1.1 (Figure 4, panels A, B). The nomenclature is based on the tree topology, reflecting the order of the nodes from the root to the tips, but it is unrelated to the sequence collection date or date of the most recent common ancestor of the lineage.

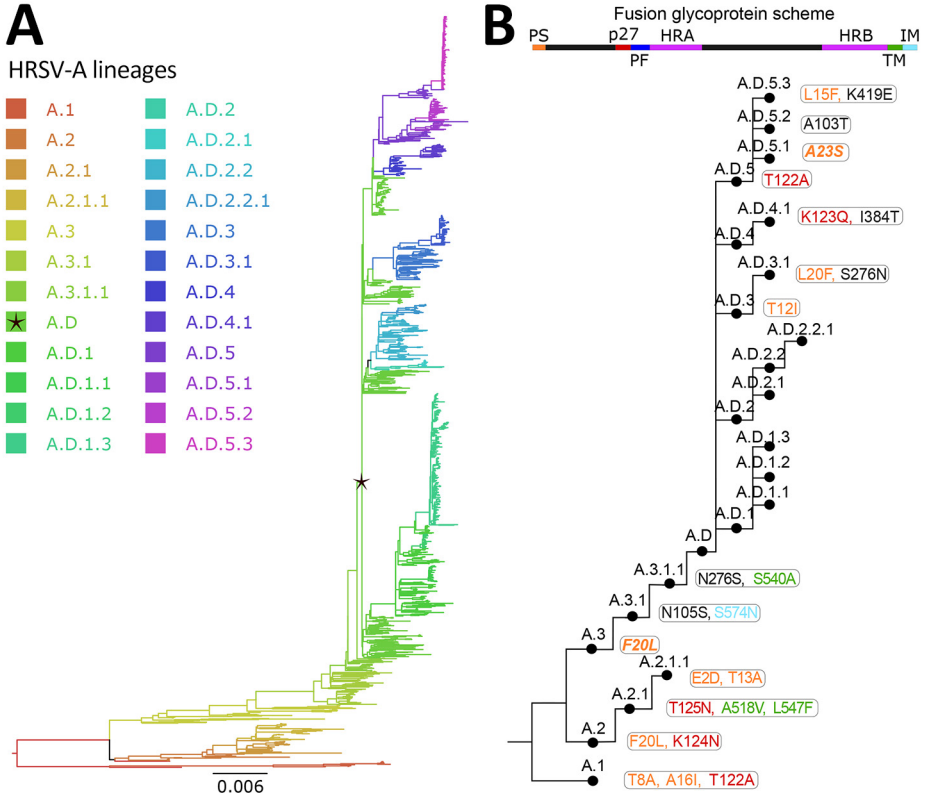


Figure 3. Human respiratory syncytial virus A lineage classification. A) HRSV-A maximum-likelihood phylogenetic tree (1,480 sequences), colored by lineage classification. Black star indicates A.D lineage, defined by the 72-nt duplication in the G gene. Scale bar indicates substitutions per site. B) Simplified scheme of the lineage designation to highlight the presence of nested lineages. The amino acid changes in the F glycoprotein are listed next to lineage name and colored according to their location in the fusion protein.

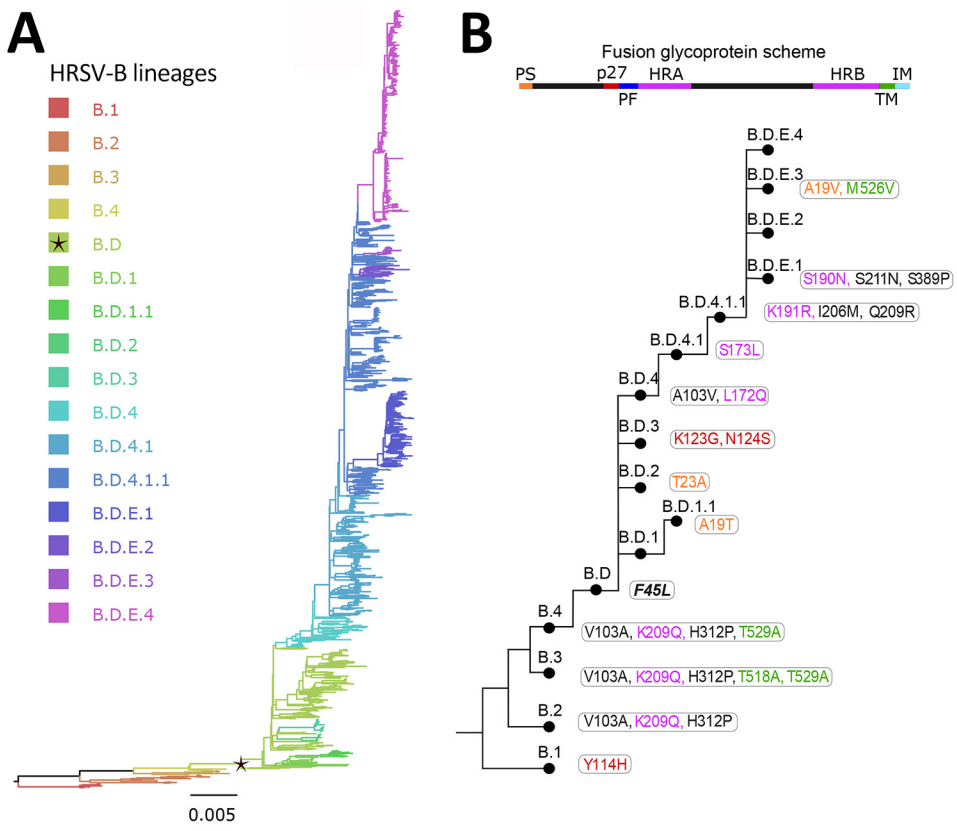


Figure 4. Human respiratory syncytial virus B lineages classification. A) HRSV-B maximum-likelihood phylogenetic tree (1,385 sequences), colored according to lineage classification. Black star indicates B.D lineage, defined by the 60-nt duplication in the G gene. Scale bar indicates substitutions per site. B) Simplified scheme of the lineage designation to highlight the presence of nested lineages. The amino acid changes in the F glycoprotein are listed next to lineage name and colored according to their location in the fusion protein.

To remain functional, a nomenclature system requires periodic updates as new lineages emerge. Therefore, we have established 2 open repositories on GitHub containing definitions of each lineage, signature mutations, and representative sequences. The repositories are available at https://github.com/rsv-lineages/lineage-designation-A and https://github.com/rsv-lineages/lineage-designation-B; they are intended to provide up-to-date definitions and serve as a platform for discussion and designation of novel lineages.

Lineages within the HRSV-A and HRSV-B Rooted Trees

We reconstructed ancestral sequences at the root of the phylogenetic trees. Although the sequences are not biologically real, they served as surrogate parental lineages during initial classification. Identifying monophyletic clusters with ≥10 sequences and ≥5 aa changes compared with the reconstructed root sequence, we defined 3 HRSV-A lineages (A.1–A.3) and 4 HRSV-B lineages (B.1–B.4). We were unable to classify 2 sequences, EPI-ISL-15771600_USA_1956 (GISAID) and MG642074_USA_1980 (GenBank), perhaps because they belong to underrepresented extinct lineages.

We further analyzed the first lineages in an iterative manner to identify nested lineages; as a result, we identified a total of 24 lineages within HRSV-A, and 16 within HRSV-B (Figures 3, 4). Close to the root of the HRSV-B tree, extinct lineages were underrepresented, comprising <10 sequences but featuring >5 distinct amino acids (B.1, B.3, B.4). Despite the low number of sequences, we included them as lineages to trace evolutionary branches that gave rise to currently circulating lineages. In addition, A.D.2 is slightly below the sequence threshold; nonetheless, we kept the lineage category to emphasize the common ancestor among A.D.2.1 and A.D.2.2.

We scrutinized the presence and absence of the duplication in the G gene across each tree. Although patterns were mostly as expected with a single historical duplication event, some genomes within the clade with the duplication in G lacked the duplication. The dispersed association of these sequences in the phylogenetic tree, rather than the monophyletic cluster we expected, suggests the virus did not lose the nucleotide duplication (Appendix 1 Figure 4). Instead, similar read length to the duplication region of certain short-read next-generation sequencing technologies potentially masked the presence of the duplication when used in the consensus genome assembly with reference sequences that do not possess the nucleotide duplication. Therefore, we recommend

using such data with quality filtered reads of a length >150 nt to avoid this problem.

Lineage-defining amino acids were present in all HRSV proteins, primarily identified within the G protein (Tables 1, 2). Also, the lineage-defining amino acids at polymerase L protein were noteworthy, contributing to the distinction of 21 of 24 HRSV-A lineages and 15 of 16 HRSV-B lineages (Tables 1, 2). Of interest, the F protein contributed to define 14 lineages in HRSV-A and 13 in HRSV-B (Figure 3, panel B; Figure 4, panel B). The G and F surface glycoproteins are likely under selection pressure from antibody-mediated immunity and exhibit a robust phylogenetic signal (*18*,*31*). Whereas the G protein displays substantial nucleotide and amino acid sequence plasticity, the F protein experiences strong negative selection, likely attributed to functional or structural constraints (*34*). For instance, the fusion peptide is the only region in F without lineage-defining amino acids (Figure 3, panel B; Figure 4, panel B). Although the low diversity of the F protein is promising for HRSV interventions, monitoring the F protein during global implementation is essential to estimate the antigenic impact of amino acid substitutions.

Using G and F Sequences with the HRSV Lineage Classification System

The main challenge for global expansion of HRSV genomics is the absence of a cost-effective, globally standardized and validated methodology for sequencing, in contrast to SARS-CoV-2 or influenza virus (*35*,*36*). In addition, limited funding and infrastructure cause some laboratories to prefer sequencing the G gene only (*37*–*39*). Although we highly recommend using complete genomes for HRSV lineage assignment to ensure the maximum accuracy of the classification and monitor the amino acid changes in all viral proteins, partial genomes covering the G and F genes can be used because overall they reproduce the topology of the HRSV tree (*17*,*18*). We do not recommend the use of smaller G gene regions such as the second hypervariable region (250-nt length at the 3′ gene end) (Figure 1) that was used historically for molecular epidemiology because previous reports showed a decreased phylogenetic signal (*17*). The use of G, F, or both genes for lineage classification should rely on phylogenetic associations with reference sequences. Of note, using only G and F genes is inadequate for defining novel lineages because of the inability to detect lineage-defining amino acids across all viral proteins. Our analysis showed minimal misclassification (1.2%) in HRSV-A and none in HRSV-B when using only the G gene (Appendix 1 Figure 5). However, the G ectodomain alone resulted in an 18.86% misclassification rate for HRSV-A and none for HRSV-B. The F gene alone had misclassification rates of 38.18% for HRSV-A and 1.23% for HRSV-B because of polytomies affecting lineage assignments within A.D.1 and A.D.5. Combining G and F gene fragments reduced misclassification to 0.07% for HRSV-A and none for HRSV-B, indicating that this approach provides optimal resolution for both subgroups (Appendix 1 Figure 5).

Prospective HRSV Lineage Assignment and Definition

Assigning sequences to the existing lineages can be automated using online tools such as NextClade (https://clades.nextstrain.org) (*40*), ReSVidex (https://cacciabue.shinyapps.io/resvidex_wg), INSaFLU (https://insaflu.insa.pt) (*41*), or UShER (https://usher.bio) (*42*). However, to define a novel lineage, we encourage users to follow our guidelines (Appendix 1), available on GitHub (https://github.com/orgs/rsv-lineages/repositories). We anticipate new lineages of HRSV-A/B will continue to emerge, and we envision updating our proposed nomenclature to incorporate new lineages. We encourage reporting of new HRSV lineages at the RGCC GitHub page as an issue within the corresponding repository for HRSV-A/B. The RGCC study group will evaluate the newly proposed lineage and update reference alignments if confirmed.

Importantly, assigning the lineage of a query sequence does not require the use of complete genomes or the absence of nucleotide ambiguities; rather, it requires a supported association within a phylogenetic clade. However, defining a new lineage requires the use of complete genomes without ambiguities, because amino acid characterization of all viral proteins is essential.

Molecular Epidemiology of HRSV with Proposed Classification

We described the HRSV molecular epidemiology including all available genomes, even those previously discarded during the dataset curation. We analyzed the seasonality of lineages using a dataset comprising 2,277 HRSV-A and 2,058 HRSV-B genomes, revealing notable co-circulation and lineage replacement over time (Figure 5). In HRSV-A, A.1 and A.2 lineages are extinct: the last detected sequences of A.1 were collected in 1995 and of A.2 in 2015. Since 2011, A.D and nested lineages continue to circulate; A.D.2.2 and A.D.4 were detected in 2013, indicating rapid divergence of the HRSV-A viruses with the 72-nt duplication in G gene. In HRSV-B, lineages B.1, B.2, B.3, and B.4 exhibited

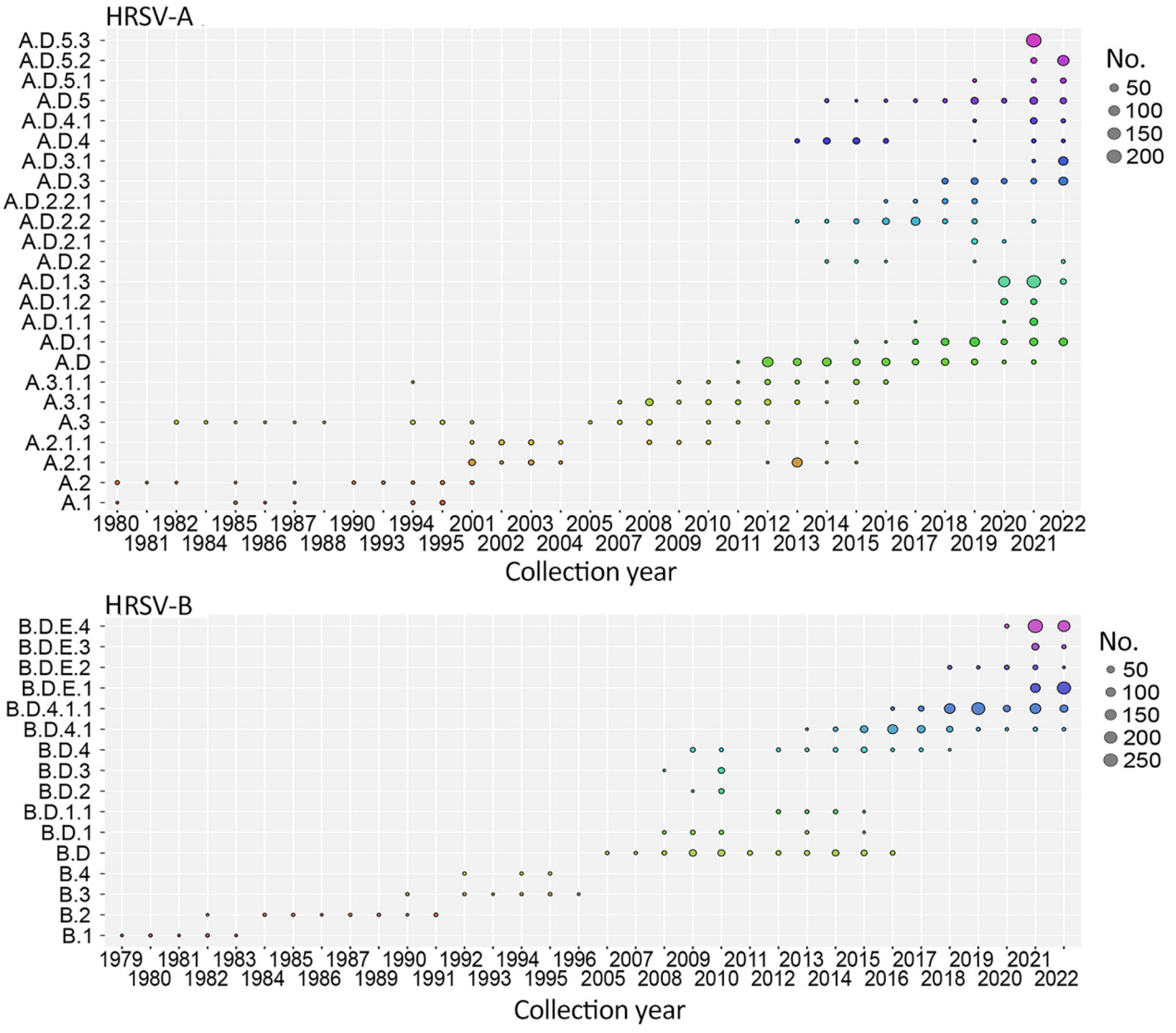


Figure 5. Temporal distribution of HRSV-A and HRSV-B lineages. A total of 2,744 HRSV-A genomes and 2,443 HRSV-B genomes available in public databases through March 2023 were included. HRSV, human respiratory syncytial virus.

strong lineage replacement (Figure 5). Although the B.D lineage with a 60-nt duplication in the G gene (B.D lineage) was detected in 1999, complete genomes became available in 2005 (*8*). By 2009, only B.D and nested lineages were detected, and since 2017, only B.D.4 and nested lineages have been observed.

HRSV lineages may have been underrepresented before the COVID-19 pandemic because of limited genomic surveillance. However, our classification system allows for updates if prepandemic genomes meeting lineage criteria are shared. Some lineages, such as A.D.3.1, A.D.5.2, and A.D.5.3 in HRSV-A and B.D.E.1 and B.D.E.3 in HRSV-B, appear to be exclusive to the postpandemic period, although most of their lineage-defining amino acid were present in parental prepandemic strains. For instance, A.D.5.2 was recognized as a distinct lineage with the emergence of the C26Y substitution in M2–2, whereas other signature amino acids were present in a 2019 parental lineage genome (GenBank accession no. MZ515825). Detection of postpandemic lineages does not contradict studies reporting no new post-pandemic genotypes because those studies relied on earlier classification systems (*43*–*46*). The possibility that these new lineages circulated before the pandemic depends on the deposition of genomes.

Some of the lineages were detected in specific countries (Appendix 1 Figure 6). For example, A.D.1 descendant lineages, A.D.5.3 and most of B.D.E.4 cases were identified in Australia or New Zealand.

Contemporary lineages such as B.D.4.1.1 and descendants B.D.E.1 and B.D.E.3, predominantly consisted of sequences from the United Kingdom. Global genomic surveillance bias presents a major confounding factor in lineage geodetection; for instance, most of the earliest lineages were detected in the United States, the principal contributor of HRSV genomes until 2007 (Appendix 1 Figures 2, 6).

Discussion

Consensus classification of HRSV below the subgroup level has been a challenge for multiple decades. Collaboratively, the HRSV molecular evolution research community, along with experts in the evolution of other respiratory viruses, have worked toward establishing a unified global classification system in the initiative HRSV Genotyping Consensus Consortium (RGCC). Our proposal categorizes HRSV-A/B sequences into lineages based on phylogenetic associations and amino acid markers, relying on complete genomes. Partial or low-quality genomes can be assigned to the existing lineages, emphasizing the robustness of this system. We developed standard guidelines for lineage definition and assignment and created online resources for updates, ensuring long-term utility. Defining a viral category below species through a phylogenetic-based classification is challenging; the system must exhibit reproducibility, balance complexity, and be updatable to capture the level of heterogeneity useful for viral surveillance. Our proposal addresses those requirements comprehensively.

HRSV is not an emerging virus; it generates annual outbreaks with co-circulation and replacement in the prevalence of its antigenic subgroups. Although some HRSV genomes were collected from clinical samples >50 years ago, the largest increase in the number of genomes has occurred since 2021. A limitation of our definition is the uncertainty of the antigenic effect of individual amino acid substitutions on lineages. Hence, whole-genome surveillance together with the study of lineage-phenotype association are essential, as observed in genetic and antigenic characterization in influenza to estimate the effectiveness of immunization (*47*). In 2023, recombinant F protein vaccines were approved; as their implementation progresses, we will learn how the vaccines affect viral evolution. We expect our unification proposal for the phylogenetic classification of HRSV to support spatiotemporal comparative lineage surveillance and detection of emerging lineages. In addition, we anticipate studies of association between lineages and the severity of HRSV disease, as well as associations of particular lineages with patients' demographic characteristics.

This article was preprinted at https://www.medrxiv.org/content/10.1101/2024.02.13.24302237v1.

Acknowledgments

We acknowledge the authors who have shared HRSV genomes on the public databases National Center for Biotechnology Information Virus, GenBank, European Nucleotide Archive, DDBJ, and GISAID EpiRSV. We thank the researchers and public health scientists who provided valuable comments during the initial stages of the RSV Genotyping Consensus Consortium's work.

Authors from second to last are listed alphabetically.

R.A.N. consults for Moderna on matter in virus evolution. N.W. has received grant funding from the Bill and Melinda Gates Foundation and Sanofi. The authors received no financial support for the research, authorship, or publication of this article.

About the Author

Dr. Goya is a postdoctoral researcher in the Department Laboratory of Medicine and Pathology at the University of Washington. Her work focuses on respiratory virus evolution and interactions with the immune system.

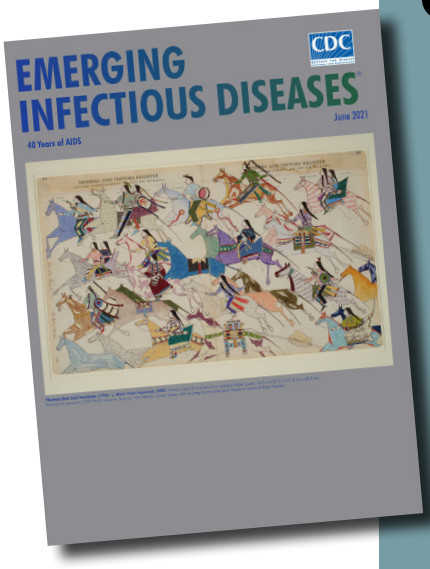
References

1. European Medicine Agency. Arexvy. 2023 [cited 2024 Jun 28]. https://www.ema.europa.eu/en/medicines/human/EPAR/arexvy
2. US Food and Drug Administration. Abrysvo. 2023 [cited 2024 Jun 28]. https://www.fda.gov/vaccines-blood-biologics/abrysvo
3. US Food and Drug Administration. Nirsevimab. 2023 [cited 2023 Nov 12]. https://www.fda.gov/news-events/press-announcements/fda-approves-new-drug-prevent-rsv-babies-and-toddlers
4. Salimi V, Viegas M, Trento A, Agoti CN, Anderson LJ, Avadhanula V, et al. Proposal for human respiratory syncytial virus nomenclature below the species level. Emerg Infect Dis. 2021;27:1–9. https://doi.org/10.3201/eid2706.204608
5. Anderson LJ, Hierholzer JC, Tsou C, Hendry RM, Fernie BF, Stone Y, et al. Antigenic characterization of respiratory syncytial virus strains with monoclonal antibodies. J Infect Dis. 1985;151:626–33. https://doi.org/10.1093/infdis/151.4.626
6. Tian D, Battles MB, Moin SM, Chen M, Modjarrad K, Kumar A, et al. Structural basis of respiratory syncytial virus subtype-dependent neutralization by an antibody targeting the fusion glycoprotein. Nat Commun. 2017;8:1877. https://doi.org/10.1038/s41467-017-01858-w
7. Peret TCT, 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;79:2221–9. https://doi.org/10.1099/0022-1317-79-9-2221
8. Trento A, Galiano M, Videla C, Carballal G, García-Barreno B, Melero JA, 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;84:3115–20. https://doi.org/10.1099/vir.0.19357-0
9. Eshaghi A, Duvvuri VR, Lai R, Nadarajah JT, Li A, Patel SN, 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:e32807. https://doi.org/10.1371/journal.pone.0032807
10. 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;82:2117–24. https://doi.org/10.1099/0022-1317-82-9-2117
11. Cui G, Zhu R, Qian Y, Deng J, Zhao L, Sun 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:e75020. https://doi.org/10.1371/journal.pone.0075020
12. Hirano E, Kobayashi M, Tsukagoshi H, Yoshida LM, Kuroda M, Noda M, et al. Molecular evolution of human respiratory syncytial virus attachment glycoprotein (G) gene of new genotype ON1 and ancestor NA1. Infect Genet Evol. 2014;28:183–91. https://doi.org/10.1016/j.meegid.2014.09.030
13. Blanc A, Delfraro A, Frabasile S, Arbiza J. Genotypes of respiratory syncytial virus group B identified in Uruguay. Arch Virol. 2005;150:603–9. https://doi.org/10.1007/s00705-004-0412-x
14. Dapat IC, Shobugawa Y, Sano Y, Saito R, Sasaki A, Suzuki Y, et al. New genotypes within respiratory syncytial virus group B genotype BA in Niigata, Japan. J Clin Microbiol. 2010;48:3423–7. https://doi.org/10.1128/JCM.00646-10
15. Shobugawa Y, Saito R, Sano Y, Zaraket H, Suzuki Y, Kumaki A, et al. Emerging genotypes of human respiratory syncytial virus subgroup A among patients in Japan. J Clin Microbiol. 2009;47:2475–82. https://doi.org/10.1128/JCM.00115-09
16. Muñoz-Escalante JC, Comas-García A, Bernal-Silva S, Robles-Espinoza CD, Gómez-Leal G, Noyola DE. Respiratory syncytial virus A genotype classification based on systematic intergenotypic and intragenotypic sequence analysis. Sci Rep. 2019;9:20097. https://doi.org/10.1038/s41598-019-56552-2
17. Goya S, Galiano M, Nauwelaers I, Trento A, Openshaw PJ, Mistchenko AS, et al. Toward unified molecular surveillance of RSV: a proposal for genotype definition. Influenza Other Respir Viruses. 2020;14:274–85. https://doi.org/10.1111/irv.12715
18. Ramaekers K, Rector A, Cuypers L, Lemey P, Keyaerts E, Van Ranst M. Towards a unified classification for human respiratory syncytial virus genotypes. Virus Evol. 2020;6:veaa052. https://doi.org/10.1093/ve/veaa052 https://doi.org/10.1093/ve/veaa052
19. Chen J, Qiu X, Avadhanula V, Shepard SS, Kim DK, Hixson J, et al. Novel and extendable genotyping system for human respiratory syncytial virus based on whole-genome sequence analysis. Influenza Other Respir Viruses. 2022;16:492–500. https://doi.org/10.1111/irv.12936
20. 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. 2002;30:3059–66. https://doi.org/10.1093/nar/gkf436
21. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 2014;30:3276–8. https://doi.org/10.1093/bioinformatics/btu531
22. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37:1530–4. https://doi.org/10.1093/molbev/msaa015
23. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol. 2018;35:518–22. https://doi.org/10.1093/molbev/msx281
24. Sagulenko P, Puller V, Neher RA. TreeTime: maximum-likelihood phylodynamic analysis. Virus Evol. 2018;4:vex042. https://doi.org/10.1093/ve/vex042
25. Huddleston J, Hadfield J, Sibley TR, Lee J, Fay K, Ilcisin M, et al. Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. J Open Source Softw. 2021;6:2906. https://doi.org/10.21105/joss.02906
26. McBroome J, de Bernardi Schneider A, Roemer C, Wolfinger MT, Hinrichs AS, O'Toole AN, et al. A framework for automated scalable designation of viral pathogen lineages from genomic data. Nat Microbiol. 2024;9,550–60. https://doi.org/10.1038/s41564-023-01587-5
27. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol. 2021;7:veab064.
28. World Health Organization/World Organisation for Animal Health/Food and Agriculture Organization (WHO/OIE/FAO) H5N1 Evolution Working Group. Revised and updated nomenclature for highly pathogenic avian influenza A (H5N1) viruses. Influenza Other Respir Viruses. 2014;8:384–8. https://doi.org/10.1111/irv.12230
29. Hassan AS, Pybus OG, Sanders EJ, Albert J, Esbjörnsson J. Defining HIV-1 transmission clusters based on sequence data. AIDS. 2017;31:1211–22. https://doi.org/10.1097/QAD.0000000000001470
30. Streng A, Goettler D, Haerlein M, Lehmann L, Ulrich K, Prifert C, et al. Spread and clinical severity of respiratory syncytial virus A genotype ON1 in Germany, 2011–2017. BMC Infect Dis. 2019;19:613. https://doi.org/10.1186/s12879-019-4266-y
31. Goya S, Lucion MF, Shilts MH, Juárez MDV, Gentile A, Mistchenko AS, et al. Evolutionary dynamics of respiratory syncytial virus in Buenos Aires: viral diversity, migration, and subgroup replacement. Virus Evol. 2023;9:vead006.
32. Liang X, Liu DH, Chen D, Guo L, Yang H, Shi YS, et al. Gradual replacement of all previously circulating respiratory syncytial virus A strain with the novel ON1 genotype in Lanzhou from 2010 to 2017. Medicine (Baltimore). 2019;98:e15542. https://doi.org/10.1097/MD.0000000000015542
33. van Niekerk S, Venter M. Replacement of previously circulating respiratory syncytial virus subtype B strains with the BA genotype in South Africa. J Virol. 2011;85:8789–97. https://doi.org/10.1128/JVI.02623-10
34. Hause AM, Henke DM, Avadhanula V, Shaw CA, Tapia LI, Piedra PA. Sequence variability of the respiratory syncytial virus (RSV) fusion gene among contemporary and historical genotypes of RSV/A and RSV/B. PLoS One. 2017; 12:e0175792. https://doi.org/10.1371/journal.pone.0175792
35. Quick J. nCoV-2019 sequencing protocol v1. 2020 Jan [cited 2023 Nov 12]. https://www.protocols.click/view/ncov-2019-sequencing-protocol-bbmuik6w
36. Zhou B, Wentworth DE. Influenza A virus molecular virology techniques. In: Kawaoka Y, Neumann G, editors. Influenza virus: methods and protocols. Totowa, NJ:

Humana Press; 2012. p. 175–92 [cited 2023 Nov 12]. https://doi.org/10.1007/978-1-61779-621-0_11
37. Dong X, Deng YM, Aziz A, Whitney P, Clark J, Harris P, et al. A simplified, amplicon-based method for whole genome sequencing of human respiratory syncytial viruses. J Clin Virol. 2023;161:105423. https://doi.org/10.1016/j.jcv.2023.105423
38. Wang L, Ng TFF, Castro CJ, Marine RL, Magaña LC, Esona M, et al. Next-generation sequencing of human respiratory syncytial virus subgroups A and B genomes. J Virol Methods. 2022;299:114335. https://doi.org/10.1016/j.jviromet.2021.114335
39. Presser LD, van den Akker WMR, Meijer A, for PROMISE investigators. Respiratory Syncytial Virus European Laboratory Network 2022 survey: need for harmonization and enhanced molecular surveillance. J Infect Dis. 2023 Aug 14;jiad341.
40. Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J Open Source Softw. 2021;6:3773. https://doi.org/10.21105/joss.03773
41. Borges V, Pinheiro M, Pechirra P, Guiomar R, Gomes JP. INSaFLU: an automated open web-based bioinformatics suite "from-reads" for influenza whole-genome-sequencing-based surveillance. Genome Med. 2018;10:46. https://doi.org/10.1186/s13073-018-0555-0
42. Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, et al. Ultrafast Sample placement on Existing tRees (USHER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nat Genet. 2021;53:809–16. https://doi.org/10.1038/s41588-021-00862-7
43. Redlberger-Fritz M, Springer DN, Aberle SW, Camp JV, Aberle JH. Respiratory syncytial virus surge in 2022 caused by lineages already present before the COVID-19 pandemic. J Med Virol. 2023;95:e28830. https://doi.org/10.1002/jmv.28830
44. Goya S, Sereewit J, Pfalmer D, Nguyen TV, Bakhash SAKM, Sobolik EB, et al. Genomic characterization of respiratory syncytial virus during 2022–23 outbreak, Washington, USA. Emerg Infect Dis. 2023;29:865–8. https://doi.org/10.3201/eid2904.221834
45. Adams G, Moreno GK, Petros BA, Uddin R, Levine Z, Kotzen B, et al. Viral lineages in the 2022 RSV surge in the United States. N Engl J Med. 2023;388:1335–7. https://doi.org/10.1056/NEJMc2216153
46. Dolores A, Stephanie G, Mercedes S NJ, Érica G, Mistchenko AS, Mariana V. RSV reemergence in Argentina since the SARS-CoV-2 pandemic. J Clin Virol. 2022;149:105126. https://doi.org/10.1016/j.jcv.2022.105126
47. van Roekel C, Poukka E, Turunen T, Nohynek H, Presser L, Meijer A, et al. Effectiveness of immunisation products against medically attended respiratory syncytial virus infection: generic protocol for a test-negative case-control study. J Infect Dis. 2023;229(Supplement_1):S92–9. https://doi.org/10.1093/infdis/jiad483 https://doi.org/10.1093/infdis/jiad483

Address for correspondence: Stephanie Goya, Department of Laboratory Medicine and Pathology, University of Washington Medical Center, 850 Republican St, Seattle, WA 98109, USA; email: sgoya@uw.edu



Originally published in June 2021

etymologia revisited

Enterocytozoon bieneusi

[ˈentərəˌsaitəˈʒu:ən bɪəˈnəʊsɪ]

From the Greek *en'tĕr-ō-sī'tōn* (intestine), *kútos* (vessel, cell), and *zō'on* (animal), and the surname Bieneus, in memory of the first infected patient whose case was reported in Haiti during 1985. *Enterocytozoon bieneusi*, a member of the wide-ranging phylum Microsporidia, is the only species of this genus known to infect humans. Microsporidia are unicellular intracellular parasites closely related to fungi, although the nature of the relationship is not clear.

E. bieneusi, a spore-forming, obligate intracellular eukaryote, was discovered during the HIV/AIDS pandemic and is the main species responsible for intestinal microsporidiosis, a lethal disease before widespread use of antiretroviral therapies. More than 500 genotypes are described, which are divided into different host-specific or zoonotic groups. This pathogen is an emerging issue in solid organ transplantation, especially in renal transplant recipients.

Sources

1. Desportes I, Le Charpentier Y, Galian A, Bernard F, Cochand-Priollet B, Lavergne A, et al. Occurrence of a new microsporidan: *Enterocytozoon bieneusi* n.g., n. sp., in the enterocytes of a human patient with AIDS. J Protozool. 1985;32:250–4. https://doi.org/10.1111/j.1550-7408.1985.tb03046.x
2. Didier ES, Weiss LM. Microsporidiosis: not just in AIDS patients. Curr Opin Infect Dis. 2011;24:490–5. https://doi.org/10.1097/QCO.0b013e32834aa152
3. Han B, Weiss LM. Microsporidia: obligate intracellular pathogens within the fungal kingdom. Microbiol Spectr. 2017;5:97–113. https://doi.org/10.1128/microbiolspec.FUNK-0018-2016
4. Moniot M, Nourrisson C, Faure C, Delbac F, Favennec L, Dalle F, et al. Assessment of a multiplex PCR for the simultaneous diagnosis of intestinal cryptosporidiosis and microsporidiosis: epidemiologic report from a French prospective study. J Mol Diagn. 2021;23:417–23. https://doi.org/10.1016/j.jmoldx.2020.12.005

https//wwwnc.cdc.gov/eid/article/27/6/et2706_article

Article DOI: http://doi.org/10.3201/eid3008.240209

Standardized Phylogenetic Classification of Human Respiratory Syncytial Virus Below the Subgroup Level

Appendix 1

The supplementary description of each section appears in the order of mention within the manuscript.

Methods

HRSV Sequences Dataset

We downloaded HRSV complete genomes available from NCBI GenBank (https://www.ncbi.nlm.nih.gov/labs/virus) and GISAID EpiRSV (https://gisaid.org/) up to March 11, 2023 using a filter for sequence length above 14,000-nt, obtained from human hosts and including the year and country of the sample collection. We categorized sequences into files based on their reported subgroup resulting in 2,744 HRSV-A and 2,443 HRSV-B genomes (Appendix Figure 1). We reserved sequences containing nucleotide ambiguities, indicative of inadequate sequencing depth, for epidemiologic analysis but we excluded them from formal lineage definition. We removed genomes with any nucleotide ambiguities using BBmap (reformat, version Jan-2021). To ensure diversity without redundancy, BBmap (dedupe, version Feb-2020) was used to remove identical sequences, preserving one representative.

We aligned the sequences with MAFFT v7.490, and we inspected and corrected the alignments with Aliview v1.28, mainly in the G gene (17,21). Furthermore, we trimmed alignment ends to encompass complete genomes from the first codon of the first gene (NS1) to the last codon of the last gene (L). We considered partial genomes if the lack of sequence was within 50-nt of the genome ends. We used RSVsurver to detect and remove genomes with

nucleotide insertions or deletions causing frameshift in any open reading frame(s) (https://rsvsurver.bii.a-star.edu.sg). Following alignment trimming, the presence of identical sequences with nucleotide differences in the trimmed region prompted another round of redundancy removal using BBmap tool, resulting in the final set of 1,538 HRSV-A and 1,387 HRSV-B genomes (Appendix Figure 1).

Phylogenetic Analysis

We constructed maximum likelihood phylogenetic trees with IQ-TREE v2.2.0, using ModelFinder to find the best nucleotide substitution model (*1*,*2*). The reliability of sequence clusters was evaluated with SH-aLRT (1,000 replicates) and UFBoot2 (10,000 replicates) (*3*). We considered monophyletic clades defining lineages when SH-aLRT value was ≥80% and UFBoot2 value was ≥90%. Phylogenetic trees were visualized with Figtree v1.4.4 and Auspice (*4*,*5*). We assessed the temporal signal with TempEst v1.5.3, and we inferred molecular-clock phylogenies with TreeTime (*6*,*7*).

We assessed recombination events with both alignment-based and phylogenetic-based methods. RDP4 software was used to detect and characterize the recombination events within the sequence alignments using the RDP, GENECONV, Maximum Chi Square and 3SEQ methods with default settings (*8*). The TreeKnit software assessed recombination based on topological differences between trees by comparison of phylogenies inferred with the 5′ and 3′ ends of the alignments (4,500-nt each, excluding the G gene) (*9*). The resulting tanglegram, available at https://github.com/rsv-lineages/Classification_proposal, was visualized in Auspice. Recombination assessment, using RDP4 software (alignment-based) and phylogenetic tree topology-based analysis, found no evidence of recombination among HRSV sequences. Consequently, no sequences were removed due to genetic recombination.

Results

Accurate Root Placement in HRSV Phylogenetic Trees

We used two approaches to define the correct phylogenetic tree root: a) the utilization of an outgroup, a conventional method for inferring the tree root using sequences known to be evolutionarily distant, and b) phylodynamic analysis, integrating temporal and phylogenetic patterns in virus evolution.

In the first approach, five HRSV-B genomes with the earliest collection dates were aligned with the HRSV-A dataset, and vice versa (Appendix 1 Figure 3 panels A, C). As anticipated, the five sequences from the alternative subgroup formed a distantly related clade, serving as the outgroup for rooting each tree. The second approach, without outgroups, involved reconstructing dated phylogenetic trees for each subgroup dataset, and the phylogenetic root automatically inferred by incorporating temporal information (Appendix Figure 3 panels B, D). Both approaches consistently identified the same root for each subgroup cluster.

Comprehensive characterization of the datasets was achieved through phylodynamic analysis. The most recent common ancestor (MRCA) was dated to 1951 for HRSV-A and 1965 for HRSV-B (Appendix Figure 3, panels E,F). Global evolutionary rates were estimated to 7.964×10^{-4} substitutions/site/year ($r^2 = 0.96$) for HRSV-A and 6.933×10^{-4} substitutions/site/year ($r^2 = 0.98$) for HRSV-B, aligning with previous reports (*10–12*).

Outlier sequences in root-to-tip plots evaluating the temporal signal may be unreliable due to, for example, sequencing errors and/or inaccurate metadata. We identified 58 outlier sequences for HRSV-A and 2 for HRSV-B that were excluded from rates estimation and lineage designation. Following the exclusion of outliers from the root-to-tip analysis, the final dataset considered for lineage designation comprised 1,480 HRSV-A genomes and 1,385 HRSV-B genomes.

HRSV Lineage Definition

Results of the phylogenetic clustering using from 1 to 10 amino acids as thresholds in an automated manner are available at https://github.com/rsv-lineages/Classification_proposal.

Use of G and F Sequences with the HRSV Lineage Classification System

We assessed the reproducibility of the classification with the G and/or F gene sequences. The phylogenetic trees obtained are available at https://github.com/rsv-lineages/Classification_proposal.

Our assessment of using this classification with solely the G gene showed minimal misclassification (1.2% error) in HRSV-A, and none in HRSV-B (Appendix Figure 5). This result underscored strong support for lineage-defining nodes. However, employing the G ectodomain alone led to a misclassification rate of 18.86% in HRSV-A due to the association of A.D.1.3 directly from A.D (outside A.D.1 clade), with no misclassification in HRSV-B.

Conversely, relying solely on the F gene resulted in a misclassification of 38.18% and 1.23% in HRSV-A and HRSV-B sequences, respectively. The high rate of HRSV-A misclassification using solely the F gene is related with the appearance of polytomies that hampered the assignation of descendant lineages within A.D.1 and A.D.5. Interestingly, when using a fragment containing both G and F genes, misclassification dropped to 0.07% in HRSV-A and remained absent in HRSV-B, suggesting that incorporating both genes provides optimal resolution for both HRSV subgroups (Appendix Figure 5).

Prospective HRSV Lineage Assignment and Definition

HRSV Lineage Assignment

The assignment of sequences to the existing lineages can be automated using online tools such as NextClade (https://clades.nextstrain.org/) (*13*), ReSVidex (https://cacciabue.shinyapps.io/resvidex/), INSaFLU (https://insaflu.insa.pt) (*14*,*15*) or UShER (https://usher.bio/) (*16*). However, for the classical approach to define the lineage of query sequences we encourage users to follow the guidelines described below:

1. Perform an alignment (for instance, with MAFFT, online version (*17*): https://mafft.cbrc.jp/alignment/server/) with the query sequences and the most recent reference alignment of the same HRSV subgroup available in the GitHub (https://github.com/rsv-lineages/lineage-designation-A and https://github.com/rsv-lineages/lineage-designation-B).
2. Visually verify that the alignment covers the entire genomic region from the first codon of NS1 gene to the codon of the L gene. If the target sequences are longer than the reference alignments, trim the ends accordingly. If the target sequences are shorter, no trimming is necessary.
3. Visually evaluate the alignment, especially around the G gene region, to detect and correct alignment artifacts.
4. Infer a maximum likelihood phylogenetic tree with software such as IQ-TREE enabling the selection of the nucleotide substitution model with ModelFinder and assessing node support with UFBoot2 and SH-alrt. Phylogenetic analysis can be

run online (http://iqtree.cibiv.univie.ac.at/) (*18*), or locally using the command line as follows (version IQ-TREE v2.2.0):

iqtree2 -s <path_to_sequences> -alrt 1000 -B 1000

5. Visualize the phylogenetic tree with software such as FigTree. Ensure that the tree is properly rooted against A.1 or B.1 lineage clades. Find the lineage of the query sequences by associating them with the reference sequences. Evaluate whether the query sequences form a monophyletic clade with statistical support (≥90% for UFBoot2 and ≥80% for SH-alrt).

If, for example, one exclusively uses G gene sequences, the lineage assignment will be based in phylogenetic association. Although all lineages contain amino acid markers in the G gene, there are amino acid- defining lineages in other genes that will not be able to be detected. When using the G gene sequence for lineage assignment, it is recommended to trim the reference alignment from GitHub to the longest G open reading frame and verify the alignment including the query sequences mainly around the duplication region. Interpretating the results when complete genomes are not used should take into consideration the intrinsic limitations.
Inadequate sequencing depth often leads to missing data in genetic sequences, typically denoted by the letter 'N'. This missing data within a sequence may impact the accuracy of its phylogenetic clade association since in maximum likelihood trees constructed with IQ-TREE, missing characters are treated similarly to nucleotide gaps (http://www.iqtree.org/doc/Frequently-Asked-Questions). Consequently, the estimation of nucleotide site-likelihood relies on the data contained in the sequences with non-gap/missing data. However, nucleotide ambiguities, which represent multiple possible bases (e.g., R to represent A or G -purine-, Y to represent C or T -pyrimidine-), are supported in a manner that all represented bases are considered to have equal likelihood. It is crucial to note that sequences containing nucleotide ambiguities, whether they are missing data (N) or represent more than one base (such as R and Y), are still used for phylogenetic classification and identification of the HRSV lineage. However, the presence of any nucleotide ambiguities can hinder the identification of a given lineage-defining amino acid. Therefore, the definition of novel lineages requires complete genomes without any nucleotide ambiguities.

We anticipate that new lineages of HRSV-A and HRSV-B will continue to emerge in future and envision our proposed nomenclature being expanded to incorporate new lineages. The detection and definition of a new lineage comprise the use of complete genomes that should adhere to the considerations outlined in this study on the phylogenetic and amino acid criteria. The recommended procedure to define a new lineage is described below:

1. Perform an alignment (for instance, with MAFFT, online version (*17*): https://mafft.cbrc.jp/alignment/server/) of the complete genomes of the potential new lineage with the reference alignment available on GitHub (https://github.com/rsv-lineages/lineage-designation-A and https://github.com/rsv-lineages/lineage-designation-B).

2. Ensure that complete genomes cover information from the first codon of the first gene (NS1) to the last codon of the last gene (L) without the presence of ambiguous nucleotides. Trim the alignment if necessary to meet the defined criteria for complete genomes. Conversely, sequences can include missing data only in the first or last 50 nt of the alignment. Visually inspect the alignment to identify and correct bioinformatics artifacts, particularly in the G gene region.

3. Infer a maximum likelihood phylogenetic tree with IQTREE, enabling the selection of the nucleotide substitution model using ModelFinder, and assessing statistical support for nodes with UFBoot2 and SH-alrt. Phylogenetic analysis can be run online (http://iqtree.cibiv.univie.ac.at/) (*18*), or by the command line as follows (IQ-TREE v2.2.0):

```
iqtree2 -s <path_to_sequences> -alrt 1000 -B 1000
```

4. Evaluate the phylogenetic tree (in IQTREE v2.2.0, the file with a ".treefile" extension) to confirm that the potential new lineage meets the phylogenetic lineage criteria. This entails a monophyletic clade with at least 10 sequences of interest, supported by statistical values ≥90% for UFBoot2 and ≥80% for SH-alrt. The potential lineage should not contain sequences from any other lineage.

5. Conduct a comparative analysis of the protein sequences of the potential lineage and its parental lineage to confirm the amino acid lineage criteria. Subset the complete genomes alignment into individual open reading frames and translate the nucleotides using the universal genetic code. Identify among all the viral proteins at least 5 aa substitutions differentiating the potential lineage from the parental one. Those amino acids should present in more than 90% of the new lineage sequences.

6. If the potential lineage meets both the phylogenetic and amino acid substitution requirements, it can be named following the lineage nomenclature, including the suffix "x" to denote its potential status. For example, A.D.1.3.1pot for a descendant of A.D.1.3 and B.D.E.2.1x for a descendant of B.D.E.2. Another example based on the current state is A.D.E.1x for a descendant of A.D.2.2.1, where E is the alias for 2.2.1.

7. Share the proposal of the new lineage on the RGCC GitHub page as an 'issue' within the corresponding repository for HRSV-A (https://github.com/rsv-lineages/lineage-designation-A) or HRSV-B (https://github.com/rsv-lineages/lineage-designation-B). The RGCC study group will evaluate the new proposed lineage, and if accepted, the reference alignments will be updated.

References

1. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37:1530–4. PubMed https://doi.org/10.1093/molbev/msaa015

2. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–9. PubMed https://doi.org/10.1038/nmeth.4285

3. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018;35:518–22. PubMed https://doi.org/10.1093/molbev/msx281

4. NextStrain. auspice.us. https://auspice.us/

5. Rambaut A. Figtree. 2023. https://github.com/rambaut/figtree

6. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. 2018;4:vex042. PubMed https://doi.org/10.1093/ve/vex042

7. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2016;2:vew007. PubMed https://doi.org/10.1093/ve/vew007

8. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol. 2015;1:vev003. PubMed https://doi.org/10.1093/ve/vev003

9. Barrat-Charlaix P, Vaughan TG, Neher RA. TreeKnit: Inferring ancestral reassortment graphs of influenza viruses. PLOS Comput Biol. 2022;18:e1010394. PubMed https://doi.org/10.1371/journal.pcbi.1010394

10. Goya S, Lucion MF, Shilts MH, Juárez MDV, Gentile A, Mistchenko AS, et al. Evolutionary dynamics of respiratory syncytial virus in Buenos Aires: viral diversity, migration, and subgroup replacement. Virus Evol. 2023;9:vead006.

11. Tan L, Lemey P, Houspie L, Viveen MC, Jansen NJG, van Loon AM, et al. Genetic variability among complete human respiratory syncytial virus subgroup A genomes: bridging molecular evolutionary dynamics and epidemiology. PLoS One. 2012;7:e51439. PubMed https://doi.org/10.1371/journal.pone.0051439

12. Di Giallonardo F, Kok J, Fernandez M, Carter I, Geoghegan JL, Dwyer DE, et al. Evolution of human respiratory syncytial virus (RSV) over multiple seasons in New South Wales, Australia. Viruses. 2018;10:476. PubMed https://doi.org/10.3390/v10090476

13. Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J Open Source Softw. 2021;6:3773. https://doi.org/10.21105/joss.03773

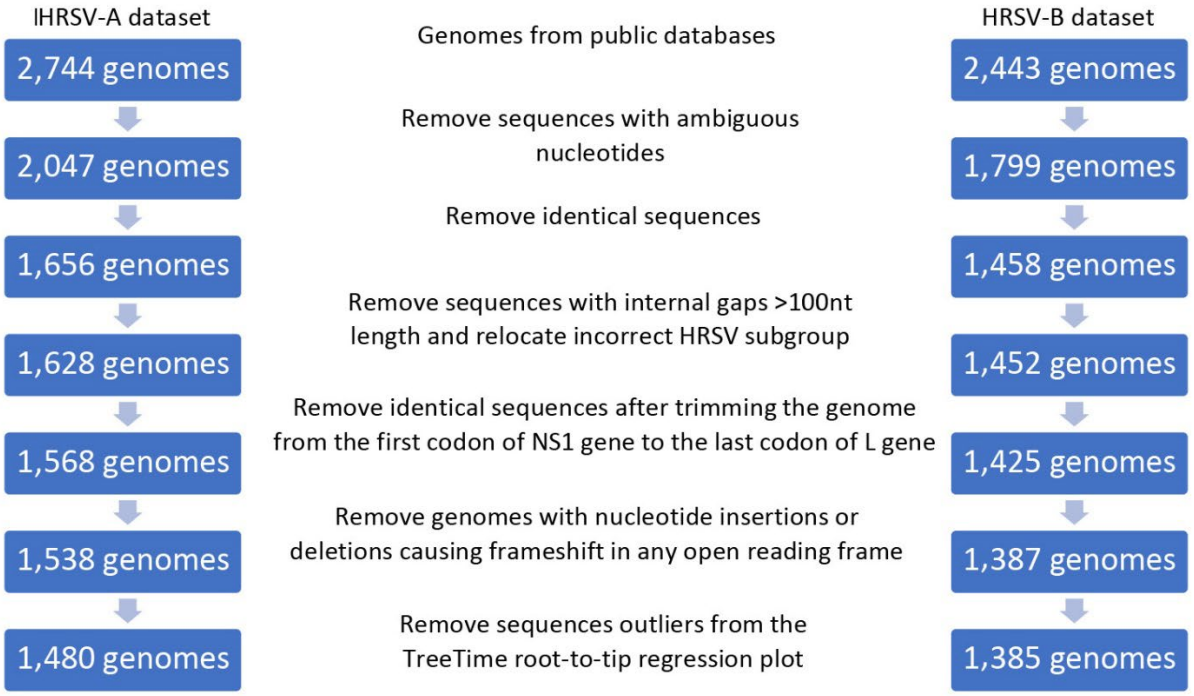
14. Borges V, Pinheiro M, Pechirra P, Guiomar R, Gomes JP. INSaFLU: an automated open web-based bioinformatics suite “from-reads” for influenza whole-genome-sequencing-based surveillance. Genome Med. 2018;10:46. PubMed https://doi.org/10.1186/s13073-018-0555-0

15. INSaFLU-TELEVIR: an open web-based bioinformatics suite for viral metagenomic detection and routine genomic surveillance. 2023. https://www.researchsquare.com

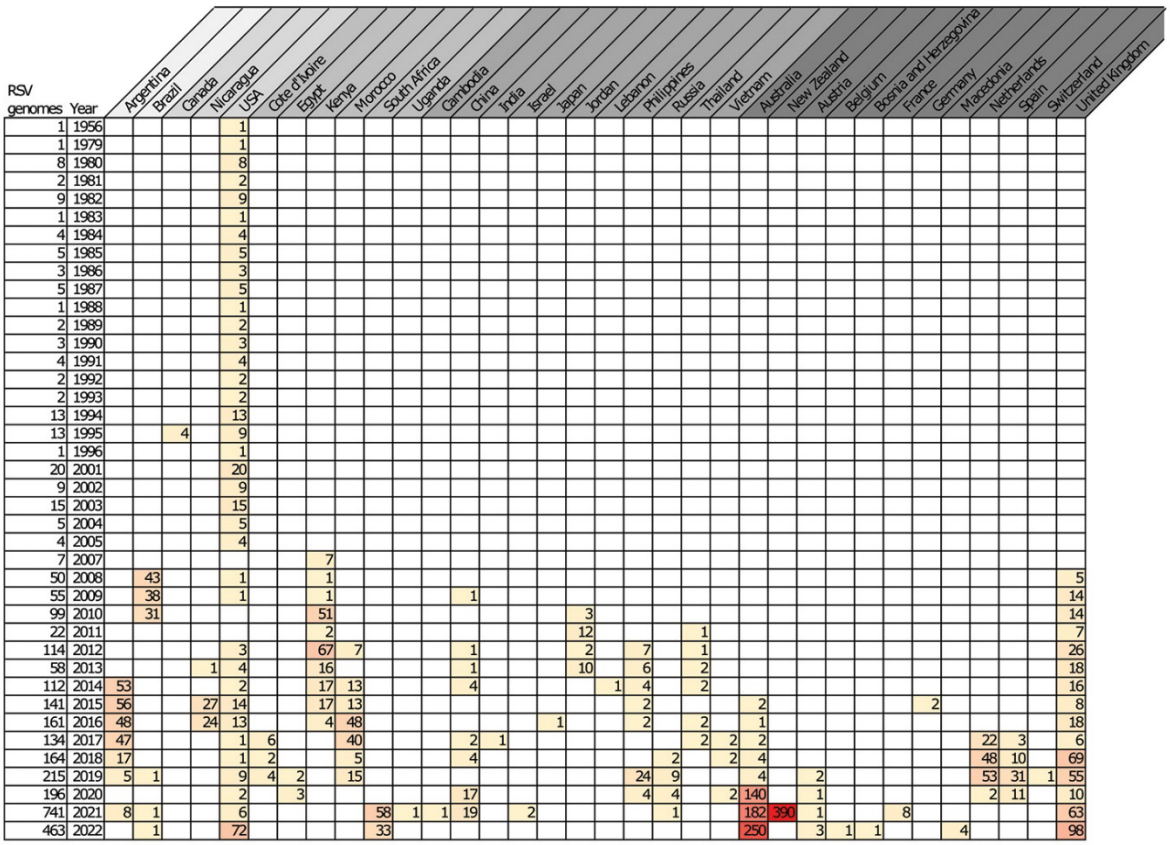
16. Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, et al. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nat Genet. 2021;53:809–16. PubMed https://doi.org/10.1038/s41588-021-00862-7

17. Rozewicki J, Li S, Amada KM, Standley DM, Katoh K. MAFFT-DASH: integrated protein sequence and structural alignment. Nucleic Acids Res. 2019;47(W1):W5–10. PubMed https://doi.org/10.1093/nar/gkz342

18. Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res. 2016;44(W1):W232-5. PubMed https://doi.org/10.1093/nar/gkw256

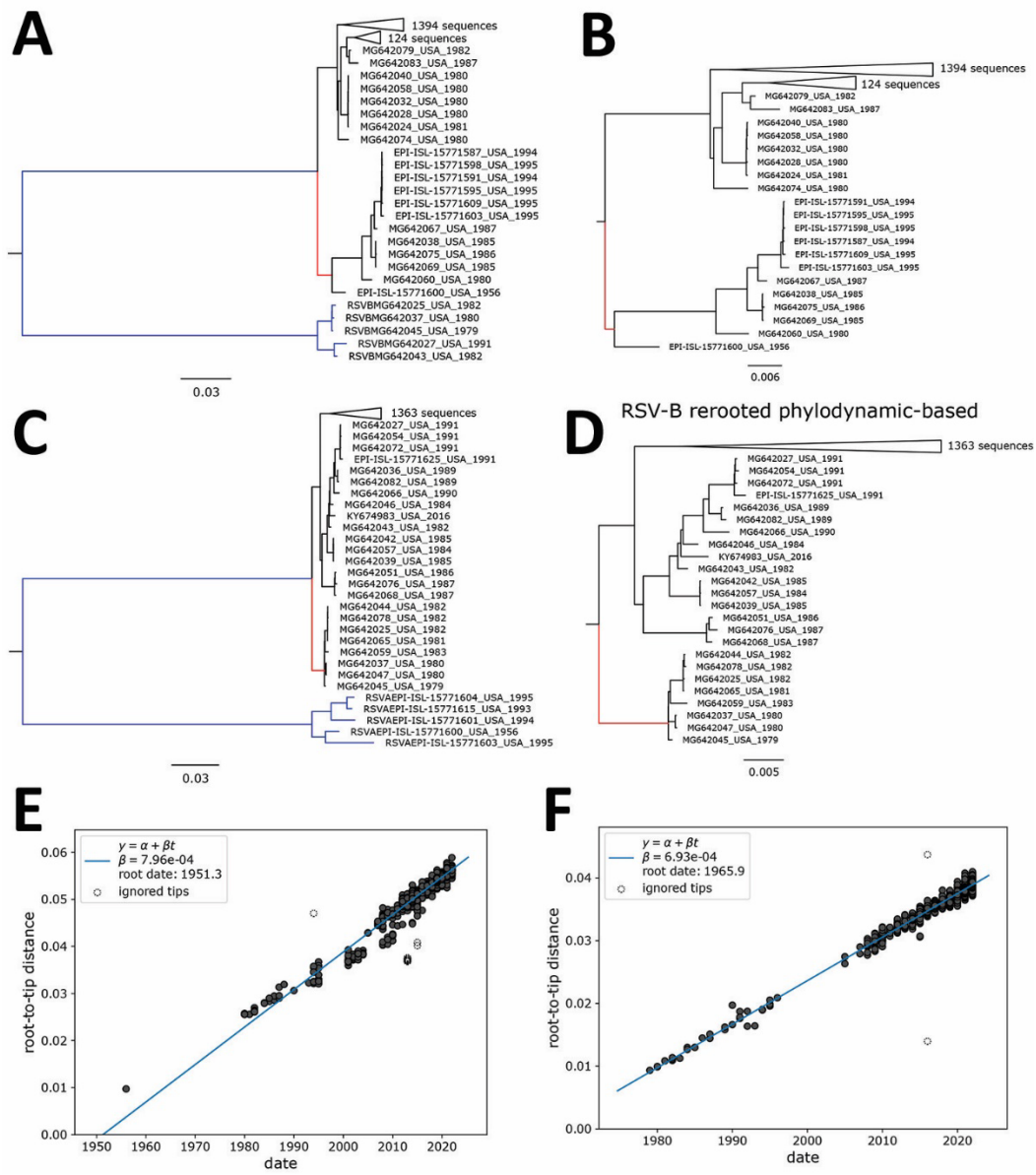


Appendix Figure 1. Dataset curation for the HRSV classification definition. For each of the filtration step during the dataset curation the number of remaining HRSV-A and HRSV-B genomes is detailed.

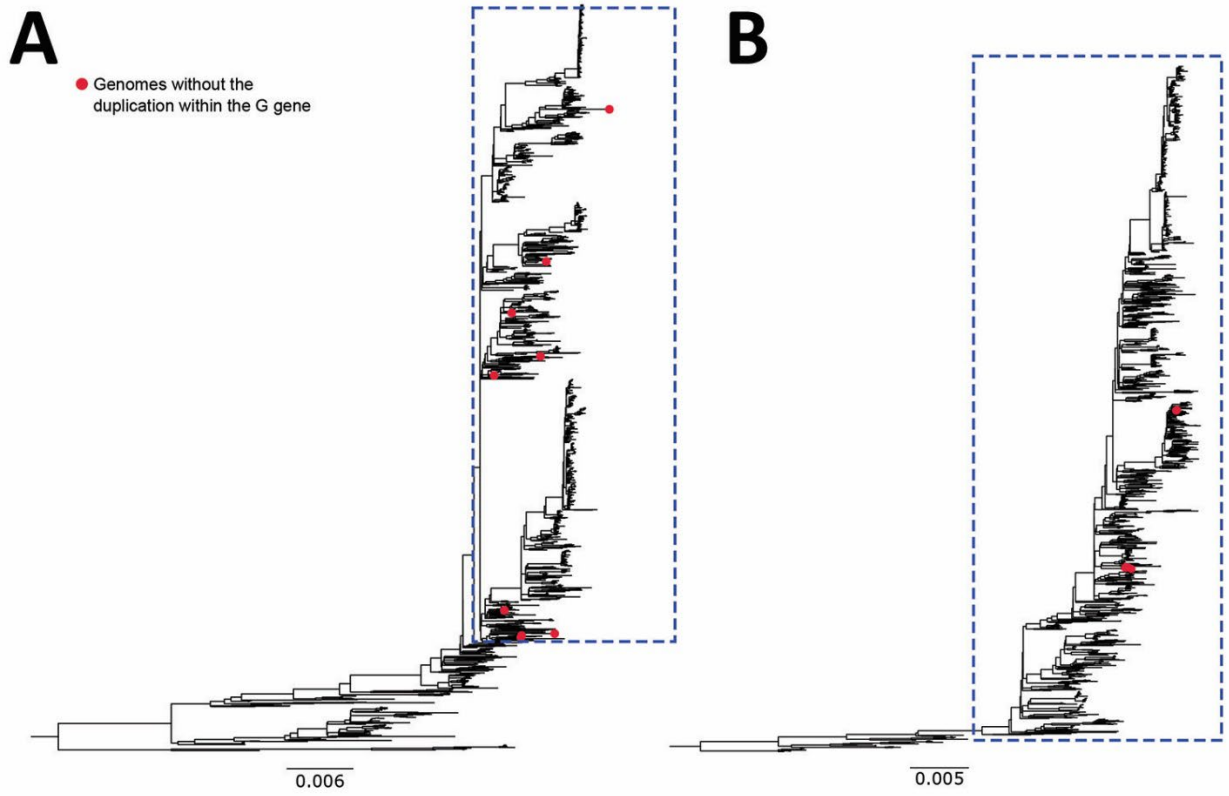


RSV genomes	Year	Argentina	Brazil	Canada	Nicaragua	USA	Cote d'Ivoire	Egypt	Kenya	Morocco	South Africa	Uganda	Cambodia	China	India	Israel	Japan	Jordan	Lebanon	Philippines	Russia	Thailand	Vietnam	Australia	New Zealand	Austria	Belgium	Bosnia and Herzegovina	France	Germany	Macedonia	Netherlands	Spain	Switzerland	United Kingdom
1	1956				1																														
1	1979				1																														
8	1980				8																														
2	1981				2																														
9	1982				9																														
1	1983				1																														
4	1984				4																														
5	1985				5																														
3	1986				3																														
5	1987				5																														
1	1988				1																														
2	1989				2																														
3	1990				3																														
4	1991				4																														
2	1992				2																														
2	1993				2																														
13	1994				13																														
13	1995			4	9																														
1	1996				1																														
20	2001				20																														
9	2002				9																														
15	2003				15																														
5	2004				5																														
4	2005				4																														
7	2007								7																										
50	2008		43		1				1																										5
55	2009		38		1				1				1																						14
99	2010		31						51								3																		14
22	2011								2								12				1														7
114	2012				3				67	7			1				2		7		1														26
58	2013			1	4				16				1				10		6		2														18
112	2014	53			2				17	13			4					1	4		2														16
141	2015	56		27	14				17	13									2				2							2					8
161	2016	48		24	13				4	48						1			2		2		1												18
134	2017	47			1	6				40			2	1								2	2	2									22	3		6
164	2018	17			1	2				5			4							2		2	4										48	10		69
215	2019	5	1		9	4	2			15									24	9			4		2									53	31	1	55
196	2020				2		3						17						4	4		2	140		1									2	11		10
741	2021	8	1		6						58	1	1	19		2					1		182	390	1				8							63
463	2022		1		72						33												250		3	1	1				4					98

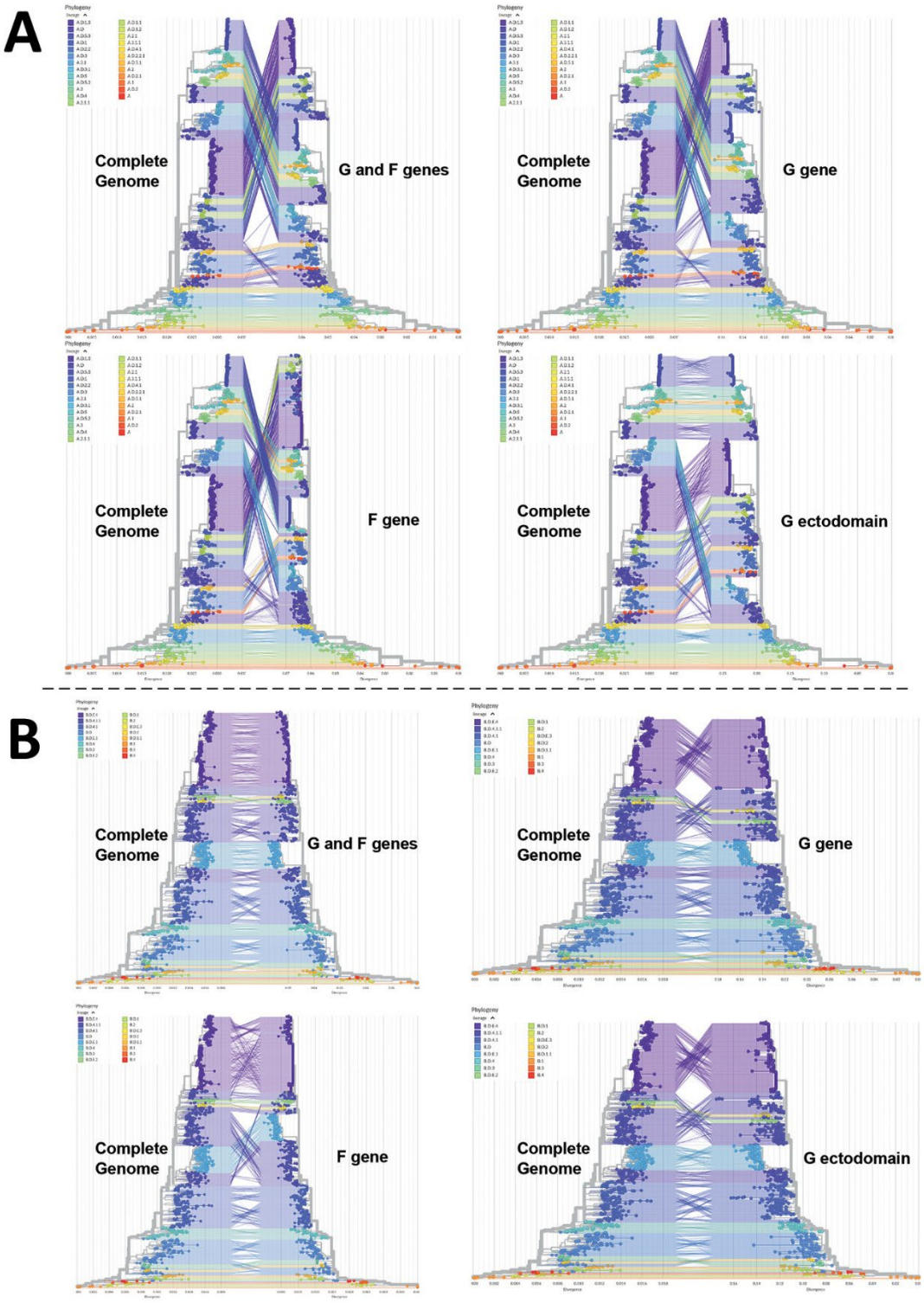
Appendix Figure 2. Characterization of the genomes used in the classification. The number of HRSV genomes per year and the country of sample collection are reported. Countries are grouped into geographic regions highlighted in gray scale. The heat map highlights the countries representation by collection year.



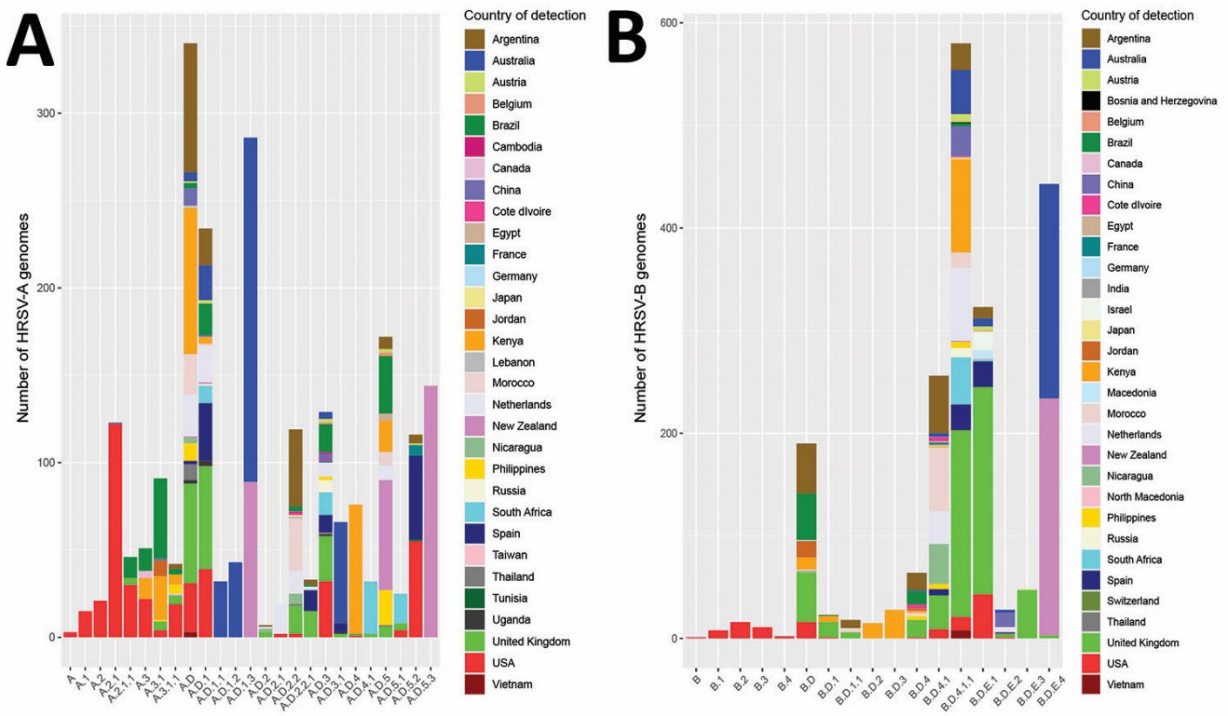
Appendix Figure 3. Rooting the HRSV phylogenetic trees. (A-B) HRSV-A maximum-likelihood (ML) phylogenetic tree rooted using an outgroup of HRSV-B sequences highlighted in blue is shown as well as the tree rooted by phylodynamic analysis. (C-D) Similarly in HRSV-B, the ML phylogenetic tree rooted using an outgroup of HRSV-A highlighted in blue and the tree rooted by phylodynamic analysis are shown. The branch indicating the common ancestor of the HRSV-A or HRSV-B sequences is indicated in red. (E-F) The genetic distances from the root to the tips of the ML tree are plotted against the year of collection. Root-to-tip mutation counts vs. sample collection date for HRSV-A and HRSV-B. Regression was plotted using to the best fitting root which minimizes the sum of the squared residuals from the regression line. The x-intercept of the regression line represents the estimated time of the most recent common ancestor of the dataset in the tree, while the gradient estimates the evolutionary rate. Outliers excluded from rate estimation are indicated.



Appendix Figure 4. Allocation of HRSV-A and HRSV-B without the duplication in the G gene. Genomes lacking the duplication of G gene in the dataset for the classification definition that were found within the clade containing the duplication of G are denoted with a red circle in the HRSV-A or HRSV-B maximum likelihood tree. The clade of A.D. or B.D. and nested lineages in HRSV-A or HRSV-B, respectively is denoted with a blue dashed line box.



Appendix Figure 5. Tanglegrams of HRSV-A and HRSV-B maximum likelihood trees constructed using complete genomes, genomic fragment comprising G and F genes, the G and F genes independently, and the G ectodomain region. The positions of the sequences in both phylogenetic trees are highlighted according to their lineage classification.



Appendix Figure 6. Geo-detection of HRSV-A and HRSV-B lineages. For each lineage, the number of genomes detected per country is reported in the stacked bar plots.