RESEARCH ARTICLE

Microevolution and phylogenomic study of Respiratory Syncytial Virus type A

Ashfaq Ahmad[1☯*], **Sidra Majaz**[1☯], **Aamir Saeed**[1☯], **Shumaila Noreen**[2], **Muhammad Abbas**[3], **Bilal Khan**[4], **Hamid Ur Rahman**[2], **Faisal Nouroz**[1], **Yingqiu Xie**[5], **Abdur Rashid**[6], **Atta Ur Rehman**[2*]

1 Department of Bioinformatics, Faculty of Natural and Computational Sciences, Hazara University, Mansehra, Khyber Pakhtunkhwa, Pakistan, **2** Department of Zoology, Faculty of Biological and Health Sciences, Hazara University, Mansehra, Khyber Pakhtunkhwa, Pakistan, **3** Department of Urology, Pakistan Institute of Medical Sciences, Islamabad, Pakistan, **4** Department of Pediatrics, Tehsil Headquarter Hospital (THQ), Dargai, Malakand, Khyber Pakhtunkhwa, Pakistan, **5** Department of Biology, School of Sciences and Humanities, Nazarbayev University, Astana, Kazakhstan, **6** Government Degree College Ara Khel, F.R Kohat, Higher Education Department, Government of Khyber Pakhtunkhwa, Kohat, Khyber Pakhtunkhwa, Pakistan

☯ These authors contributed equally to this work.
* ashfaqahmad82@hotmail.com, ashfaq.binfo@hu.edu.pk (AA); atta.rehman@hu.edu.pk (AUR)

OPEN ACCESS

Citation: Ahmad A, Majaz S, Saeed A, Noreen S, Abbas M, Khan B, et al. (2025) Microevolution and phylogenomic study of Respiratory Syncytial Virus type A. PLoS ONE 20(2): e0319437. https://doi.org/10.1371/journal.pone.0319437

Editor: Ahmed S. Abdel-Moneim, Taif University, SAUDI ARABIA

Received: September 16, 2024

Accepted: February 2, 2025

Published: February 25, 2025

Data availability statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Communal respiratory syncytial virus (RSV) causes mild to severe illnesses, predominantly in older adults, or people with certain chronic medical conditions, and in children. Symptoms may include rhinorrhea, cough, fever, and dyspnea. In most cases, the infection is mild and resolves on its own, but in some cases, it can lead to more serious illness such as bronchiolitis or pneumonia. The RSV genome codes for ten proteins, NS1, NS2, N, P, M, SH, G, F, M2 and L. We aimed to identify the RSV geographical transmission pattern based on parsimony and investigate hotspot regions across the complete RSV genomes. We employed Viral Evolutionary Network Analysis System on full-length available RSV genomes and with HyPhy for elucidating type of selection pressure. These results indicated that RSV strains circulating in South and North America are not mixed to the European samples, however, genomes reported from Australia are the direct decedents of European samples. Samples reported from the United Kingdom exhibited significant diversity, spanning almost every cluster. This report provides a complete mutational analysis of all the individual RSV genes, and particularly the 31 hotspot substituting regions circulating across the globe in RSV type A samples. Further, protein G and L displayed higher level of codons experienced positive selection. This analysis of RSV type A highlights mutational frequencies across the whole genome, offering valuable insights for epidemiological control and drug development.

Introduction

Viruses have long been a significant threat to humanity and have recently garnered increased attention due to their potential to cause large-scale pandemics, as evidenced from COVID-19 [1]. Respiratory syncytial virus (RSV), also known as human respiratory syncytial virus (hRSV) contributes to the infection of respiratory tract [2]. Initially, RSV was isolated from chimpanzee in 1956 and was also simultaneously recovered from human infants with severe lower tract respiratory disease [3]. In clinical manifestations, RSV is linked to mild upper respiratory tract illness (URTI) or otitis media to severe and potentially life-threatening lower respiratory tract involvement (LRTI). The most prevalent form of LRTI in RSV-infected newborns is bronchiolitis, however there are reports indicating pneumonia and croup. Approximately, in infants and young children ~15–50% of the lower airways were found affected in primary RSV infection that results in hospitalization and higher mortality [4,5]. Apart from supportive care like fluid intake and rest, so far there is no specific treatment for RSV infection [6].

RSV is a non-segmented negative-sense single-stranded enveloped RNA virus that belongs to the family of *Paramyxoviridae*, genus *Pneumovirus*, and subfamily *Pneumovirinae*. The complete genome of RSV encodes ten proteins, i.e., NS1, NS2, N, P, M, SH, G, F, M2 and L. Among them, M and M2 are envelope proteins, while the fusion (F) and G are glycoproteins, and SH is a small hydrophobic protein. Besides, five structural and non-structural proteins are coded by the RSV genome includingthe large (L) protein, phosphoprotein (P), nucleocapsid (N), and non-structural proteins 1 and 2 (NS1 and NS2). Among them, protein G and F are important in host cell attachment, fusion and cellular entry [5,7–9].

The RSV has been classified into two subtypes A and B which further includes strains like GA1 - GA7, SAA1, NA1 - NA4 and ON1 (RSV-A), and GB1 - GB4, SAB1 - SAB4, URU1 - URU2, BA1 - BA10, BA - C and THB (RSV-B) [10,11]. There are reports discussing RSV mutations in different lineages [12–14], however we did not find extensive analysis for the detection of hotspot regions across the whole RSV genome. Here we are reporting for the first time all the mutational frequencies, hotspot regions and transmission of RSV subtype A. These analyses highlight a wider view of RSV transmission across different geological zones that could aid in predicting the oncoming pandemic and vaccine development. Besides transmission, this report provides all the observed mutations in genome coding regions and particularly the hotspot nucleotide sites prone to mutations in individual genes of RSV.

Methodology

All the sequences used in these analyses were collected from the NCBI Virus database [15], where we selected taxonomic identification or taxid 208893, and spotted 11956 nucleotide sequences. Specific filters were applied for genome completeness, and only those genomes were considered, which were present in with sequence tag of complete, and thus the final dataset we found contained 871 sequences. The whole dataset was exported to MAFFT for whole genome alignment [16]. Next, the aligned dataset was fed to the viral genome evolutionary analysis system (VENAS) for further analyses [17].

Calculation for effective parsimony-informative site (ePIS) and network construction

To calculate ePIS, we followed a rule that the site will be considered parsimony-informative if it contains a minimum of two types of nucleotides in the aligned data, and at least two sites of them should occur with a minimum nucleotide frequency of two. Besides, the ePIS were considered effective only in the case that the site must contain unambiguous bases ≥80% of the total genomes. Keeping the above rules, 871 genomes were automatically reduced to 474 genomes were found satisfactory, and thus all the remaining analyses were carried out on the dataset containing 474 genomes. The derived ePIS results were used to classify all the genomes in haplotypes, and for this reason sequences containing similar ePIS were grouped onto the similar nodes and vice versa. All the haplotypes were next visualized by Gephi, where further haplotype networking was performed, i.e., community detection, graph rendering followed by visual inspection.

Mutational analysis

All the mutational analyses for individual genes were performed by NGS analysis package BioAider [18]. Nucleotide sequences for all the RSV individual genes were extracted from the aligned dataset of 474 genes. To classify the nature of identified mutations, each gene of RSV was handled separately, and the gaps were effectively removed. Next, we re-aligned these separated genes datasets through codon alignment tool, a BioAider functionality. Codon based alignment of each gene dataset was further analyzed for mutations and its classification. The complete reports for individual genes mutations can be found in S1 Table. To identify hotspot regions across the coding regions of the genomes, we applied a criterion that focused solely on non-synonymous substitutions. These substitutions were included only if they resulted in changes to amino acid properties and were observed in more than 200 samples. The auxiliary art work was drawn by illustrator for the visualization of biological sequences (IBS) and Paint.net [19].

Evolutionary analysis using HyPhy SLAC

To investigate the evolutionary dynamics of the RSV genes, we performed a Sitewise Likelihood Ancestral Counting (SLAC) analysis using the HyPhy software suite (version 2.5.0) [20]. SLAC is a robust method for detecting site-specific selection pressures by comparing the rates of nonsynonymous (dN) and synonymous (dS) substitutions across a codon alignment. An alignment in FASTA and tree files in Newick was prepared in MAFFT for each gene. The tree was constructed using a maximum likelihood approach with appropriate substitution models. The SLAC analyses were conducted and results were retrievd in. JSON format, which includes detailed statistics for each codon site, such as dN, dS, ω, and p-values for selection tests.

The SLAC algorithm calculates the dN/dS ratio for each codon site in the alignment. Sites with $\omega > 1$ are indicative of positive selection, while sites with $\omega < 1$ suggest purifying selection. Sites with $\omega \approx 1$ are evolving neutrally.

Nucleotide diversity, Shannon entropy and Tiajima's D calculations

To analyze the evolutionary dynamics of the RSV genes, we employed a computational pipeline implemented in Python where we employed libraries from biopython. The alignment files in FASTA format were processed to calculate three key evolutionary metrics: Shannon entropy, nucleotide diversity (π), and Tajima's D. Shannon entropy was computed for each position in the alignment to quantify site-specific variability, while nucleotide diversity provided a measure of average genetic variation across the entire sequence. Tajima's D was calculated to assess deviations from neutral evolution, with negative values indicating potential population expansion or positive selection, and positive values suggesting balancing selection or population structure. The results, including position-specific entropy, nucleotide diversity, and Tajima's D, were compiled into a CSV file for further analysis.

Results

We retrieved the complete nucleotide sequence dataset for RSV type A (taxid. 1439707) from the NCBI Virus database. By December 2022, it contained 1166 nucleotide records, including 871 complete genomes. After initial filtration and name tagging that includes genome ID, reported year and country, we applied viral genome evolutionary analysis system (VENAS) [17] and analyzed the evolutionary relationship between different genomes or RSV strains. Among them the earlier reported genome U39662.1 in 1997 was considered as a reference. Based on the effective parsimony informative sites (ePIS), and removal of redundant genome sequences, VENAS picked 474 genomes for further analyses. The final genome dataset contains sequences from the USA (2014, 2017, 2019, and 2021), Brazil (2021), Kenya (2021), Philippines (2017), Jordan (2018), Thailand (2019), Netherland (2021), Australia (2020, and 2022), China (2018), UK (2021), Spain (2021), Germany (2020), and four Unknown sequences (2022). In total our dataset 96 complete genomes samples from the USA, 26 (Spain), 38 (Brazil), 44 (Australia), 50 (UK), 62 (Netherland), 08 (Thailand), and fewer were found from other countries.

RSV transmission distribution on the country's scale

To trace the transmission routes, and contribution, we mapped 474 RSV genomes on the viral evolution network. Our data indicated individual clusters such as cluster of the USA 2014 sequences, connected to cluster of the USA 2017, which further connected the Brazilian cluster 2021. In contrast, individual sequences from Jordan 2018 and the Philippines 2017 were located at the boundaries of the Brazilian cluster, without forming distinct clusters of their own. Our analysis revealed that samples from North and South America formed distinct clusters, separate from those of the European samples. However, genomes reported from the USA 2019 were found mixed with genomes reported from Europe (Fig 1A).

To understand the transmission pattern, we used a network modularity function and clustered the whole network agnostic of country and date of collection. We retrieved 12 clades (clade A to clade L). Interestingly, our data predicted a unique pattern, for instance terminal clade A, containing purely genomes reported from the USA in the year 2014, which connected clade B retaining genomes reported from the USA in the year 2014 and 2017. Further, clade C, a direct descendent of clade B, contains all the genomes reported in 2017, except one sequence from the USA 2021. Apart from few genomes reported from the USA 2017, Philippines 2017 and Jordan 2018, clade D contain > 95% of the genomes reported from Brazil 2021. To our interest the smaller clade E, contains representation of almost all the previous clades and connects the bigger node clade F. All the previous clades did not show a single genome reported from the Europe therefore, for better understanding, we call clade F, a European clade, as it is the first with European representation that also gave birth to diverse nodes reported from European countries (Fig 1C).

The European node (clade F), is tetra-furcated into clade G, H, I and L. Among them, clade H contained samples of Nederland, Spain and the United Kingdom reported in 2021 while clade G was found enriched with all the sequences reported from Nederland 2021. Likewise, clade L is formed from the genomes reported in the United Kingdom 2021, the USA 2019, and two genomes from Spain 2021. Interestingly, clade L emerged from the European cluster through genomes reported from Philippines 2017 and the USA 2019. Finally, Clade I that contains genomes from Spain, Nederland and the United Kingdom 2021 which is bifurcated to clade J and K, where clade J contains samples from the United Kingdom 2021 that leads to Australia 2022 and the clade K only contains all the genomes reported from Australia reported in 2022. All the samples are mapped with country, year and transmission wise in (Fig 1B and 1C).

Collectively, these analyses indicate the global prevalence and the presence of different RSV type A strains. For instance, genomes reported from Australia in 2022 were gathered in two different clusters emerging from the European genomes, suggesting the presence of two different strains. However, genomes reported from Australia in 2020, are lying distantly for those reported in 2022. Likewise, the genomes from the UK reported in 2021, were found almost everywhere with European genomes, depicting the possibility of numerous RSV strains. Genomes from Spain and the Netherlands were also found in 2 and 3 different nodes, highlighting the circulation of more than one strain in that country.

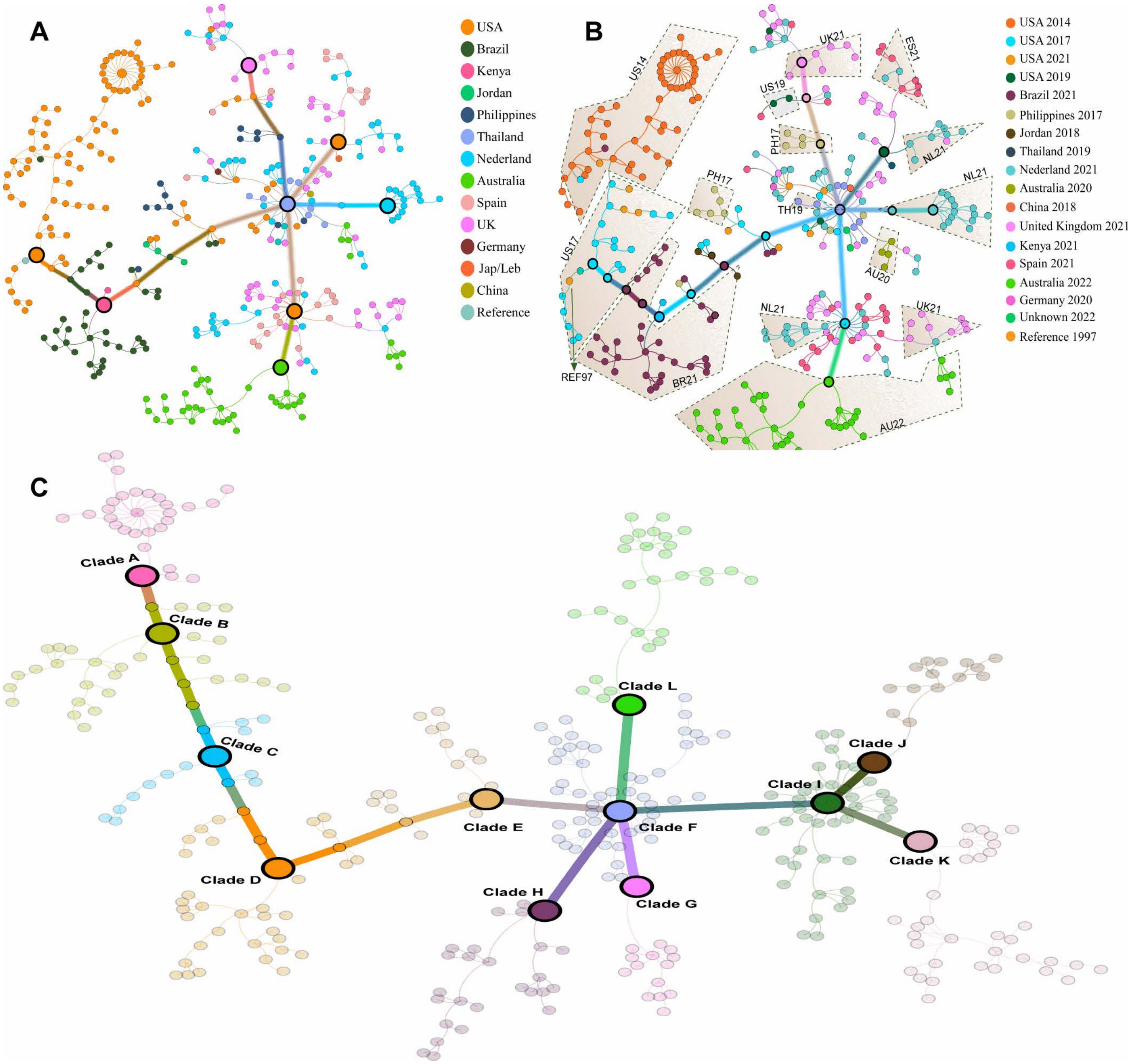

Fig 1. Transmission distribution of the RSV genomes. (A), Country wise network distribution of the RSV genomes derived through effective parsimony informative sites (ePIS). (B), Country and year wise network of complete RSV genomes. (C), calculated transmission patterns via community detection by modularity approach.

https://doi.org/10.1371/journal.pone.0319437.g001

Mutation and substitution frequencies of synonymous and non-synonymous sites in RSV genomes

According to the NCBI records, RSV contains ten to eleven protein coding genes, i.e., non-structural protein 1 and 2 (NS1 and NS2), nucleoprotein (N), phosphoprotein (P), matrix 1 protein (M), small hydrophobic protein (SH), attachment protein (G), fusion protein (F), matrix 2 protein (M2), and polymerase (L). We calculated substitution observed in RSV genomes for all coding sequences (CDS) of all ten genes, particularly synonymous and non-synonymous substitutions. All the substitutions were accessed and calculated against reference strain U39662.1. A total of 6257 (45%) sites were observed in substitutions, and among them 2099 (15.1%), 3027 (21.7%), 611 (4.3%), and 473 (3.4%) were synonymous, non-synonymous, both and terminations, respectively. Among the non-synonymous substitutions, 1442 (10%) sites were found to have changes in amino acid properties. The highest termination substitutions were found in L followed by N and G proteins (Table 1).

In protein L and G, we also observed higher and similar non-synonymous mutation frequency compared to all other genes of RSV. However, F and P proteins have relatively shown a higher number of synonymous substitutions than non-synonymous. Complete details for substitutions and substitution type for all individual genes can be accessed in form (S1 Table).

To evaluate the overall substitution frequency of the mutated sites, present in all ten CDSs, we binned the substitution frequency into seven different groups (G1 to G7) and depicted the frequency distribution of 5126 substituted sites (3027 non-synonymous and 2097 synonymous) of the 474 sequenced genomes. The group number on the X-axis indicates the number of strains participating in a substitution event at a particular site, whereas the Y-axis shows the number of substitutions in a respective CDS (Fig 2).

Our results indicated that apart from the initial groups, non-synonymous mutations were found relatively higher than the distribution of synonymous mutations. Comparing all the CDSs, the CDS of G protein showed the highest number of non-synonymous mutations where 53 mutations were found in G4 - G7, meaning the participation of 200 samples. We have also observed that CDSs of NS1, NS2, SH, and M proteins offered relatively lower sites for non-synonymous mutations in the majority of the sequenced RSV strains. The distribution of substitution frequency of each codon in each gene can be found in the supplementary S1 Table.

Overall, these analyses provide complete mutational information of all ten RSV's CDSs. Such key features can also be used to assess the evolutionary pressure on selected sites or response to therapeutic agents.

Effects of mutation and evolutionary dynamics

The investigation of evolutionary metrics such as nucleotide diversity (π), Tajima's D, Shannon entropy, and dN/dS is critical for elucidating the mechanisms underlying genetic

Table 1. Statistics of all types of substitutions observed in individual genes of RSV.

Gene	Length (CDS)	Proportion of Mutation sites	S(p)	N(p)	S-N(p)	Termination(p)
NS1	420	105 (25.0%)	79 (17.3%)	29 (6.9%)	3 (0.7%)	0
NS2	503	110 (21.8)	75 (14.9%)	30 (5.9%)	5 (0.9%)	0
N	1176	554 (47.1%)	200 (17.0%)	253 (21.5%)	67 (5.6%)	34 (2.8%)
P	726	214 (29.4%)	138 (19.0%)	67 (9.2%)	9 (1.2%)	0
M	771	177 (22.9%)	138 (17.8%)	33 (4.2%)	6 (0.7%)	0
SH	195	63 (32.3%)	33 (16.9%)	25 (12.8%)	4 (2.0%)	1 (0.5%)
G	897	496 (55.2%)	162 (18.0%)	298 (33.2%)	28 (3.1%)	8 (0.8%)
F	1725	468 (27.1%)	349 (20.2%)	106 (6.1%)	13 (0.7%)	
M2	585	150 (25.6%)	94 (16.0%)	47 (8.0%)	8 (1.3%)	1 (0.1%)
L	6498	3920 (60.32%)	831 (12.7%)	2139 (32.9%)	468 (7.2)	432 (6.6%)
Complete	13892	6257 (45.0%)	2099 (15.1%)	3027 (21.7%)	611 (4.3%)	476 (3.4%)

https://doi.org/10.1371/journal.pone.0319437.t001

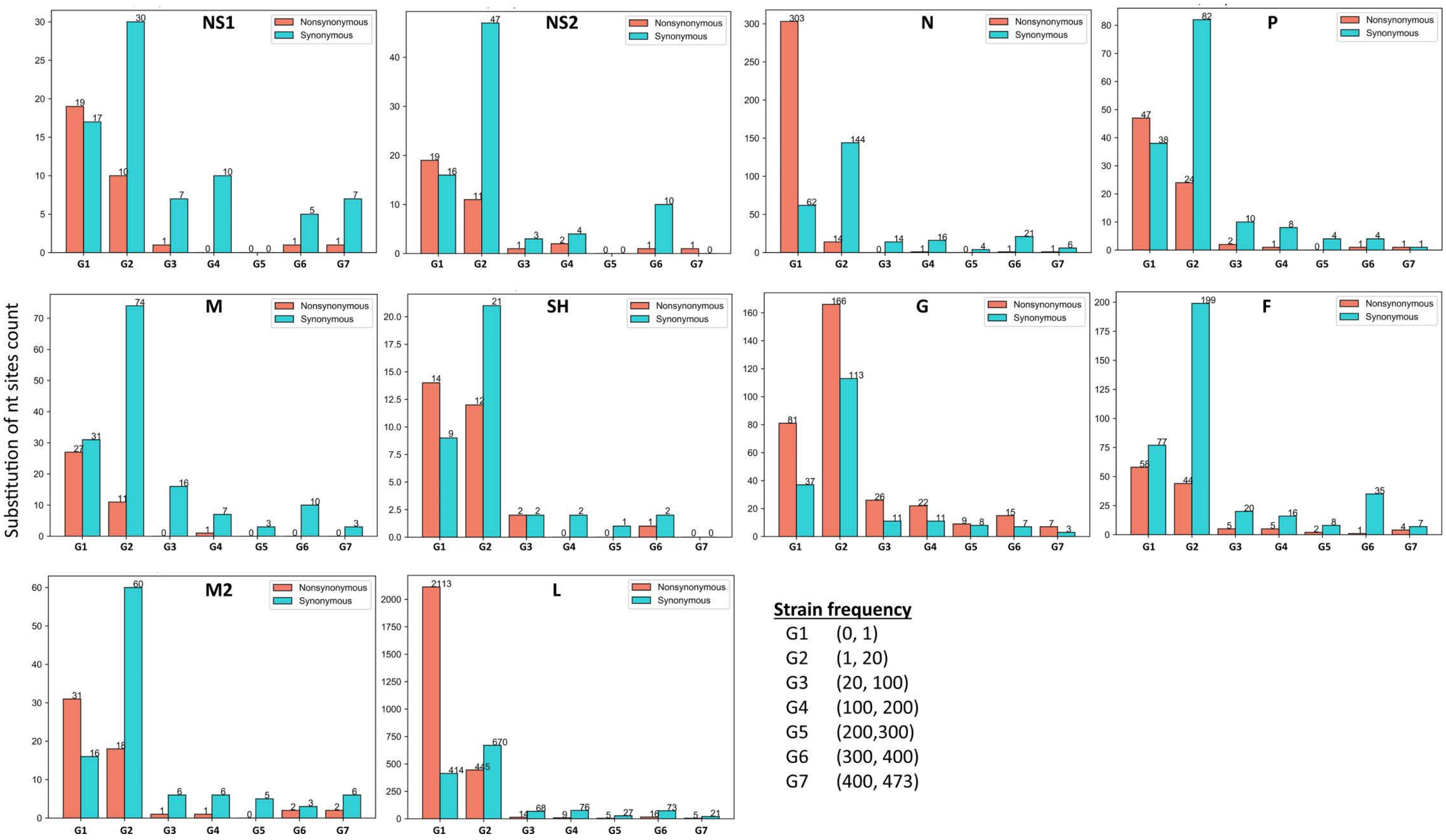

Fig 2. Overall substitution frequencies of all coding genes coded by the RSV genome. The Frequencies of only synonymous and non-synonymous substitutions are shown here. Y-axis represents the number of substitutions, and the X-axis depicts the number of genomes or samples participated.

https://doi.org/10.1371/journal.pone.0319437.g002

variation, selective pressures, and evolutionary dynamics. Each of these metrics offers distinct insights into the evolutionary processes shaping viral genomes. Our findings revealed significantly elevated nucleotide diversity (π = 0.049) in the RSV G glycoprotein gene compared to other genes, which exhibited values ranging from π = 0.017 to 0.022. This observation was further corroborated by Shannon entropy analysis, which demonstrated an average entropy value of 0.129 for the G glycoprotein gene, indicating higher variability across the RSV genome. In contrast, the remaining genes displayed average entropy values within the range of 0.0144 to 0.0160, suggesting relatively conserved regions.

Subsequent analysis using Tajima's D indicated negative values across all RSV genes, consistent with the presence of subpopulations. This outcome was anticipated, as the data set comprised samples from diverse geographical regions, reflecting population structure and potential demographic influences (Fig 3).

To assess selective pressures acting on individual genes, we computed the dN/dS ratio, which quantifies the rate of nonsynonymous to synonymous substitutions. Our results identified codons under positive selection in nearly all genes, with the G and L genes exhibiting particularly strong selective pressures. Notably, codon 115, 286, and 290 in the G gene, along with codon 1182 in the L gene, displayed dN/dS values exceeding value of 50 (S2 Table and S1 Fig), highlighting their potential roles in adaptive evolution.

In summary, these comprehensive analyses underscore the pivotal role of the G glycoprotein gene in driving RSV evolution and transmission. The elevated nucleotide diversity,

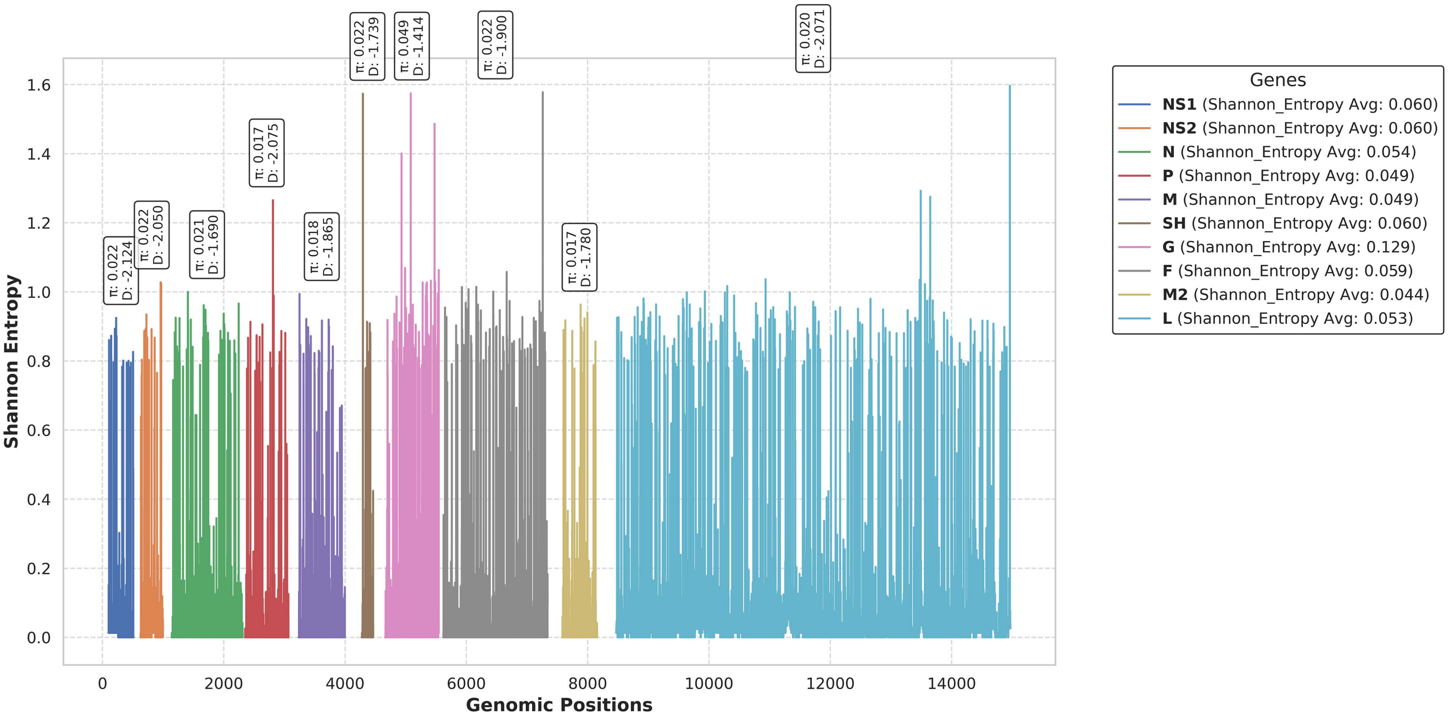

Fig 3. Statistical analysis of RSV evolution. The plot illustrates the Shannon Entropy values for each gene in the RSV genome, with genomic positions on the X-axis and entropy values on the Y-axis. Each gene is represented by a solid line and different color, and the corresponding Nucleotide Diversity (π) and Tajima's D values are displayed vertically above each gene's plot. Genomic boundaries for every gene are used from the reference RSV genome NC_001803.1. The legend, positioned outside the figure, includes the gene names in bold and their average Shannon Entropy values. The plot highlights the variability and evolutionary dynamics across the RSV genome.

https://doi.org/10.1371/journal.pone.0319437.g003

Shannon entropy, and positive selection observed in this gene suggest its significant contribution to viral adaptation and host interactions.

Mutation frequencies in non-coding regions of RSV genomes

RNA viruses, such as Respiratory Syncytial Virus (RSV), are characterized by high mutation rates, which play a critical role in their evolutionary adaptability and survival [21]. Leveraging the RSV reference genome (NC_001803), we investigated the mutational patterns within non-coding regions, with a particular focus on RNA-coding segments. According to the genomic annotations of the reference sequence, three miscellaneous RNA (misc_RNA) genes were identified, corresponding to gene IDs 1724895, 1494476, and 1724894. These genes span nucleotide positions 1 to 44 (1724895), 8460 to 8527 (1494476), and 15038 to 15191 (1724894), respectively.

Our analysis revealed a higher mutational propensity in the downstream-located gene 1724894 compared to genes 1724895 and 1494476. In contrast, the mutation frequencies in the extreme upstream and downstream regions were relatively low, ranging from 2% to 8%. Notably, four specific positions within gene 1494476; A8484G, T8485C, A8509C, and G8512T exhibited significantly higher mutation prevalence, with frequencies exceeding 30% (Fig 4). These findings highlight distinct mutational hotspots within the RSV genome, particularly in the non-coding RNA regions, which may contribute to the virus's adaptive mechanisms.

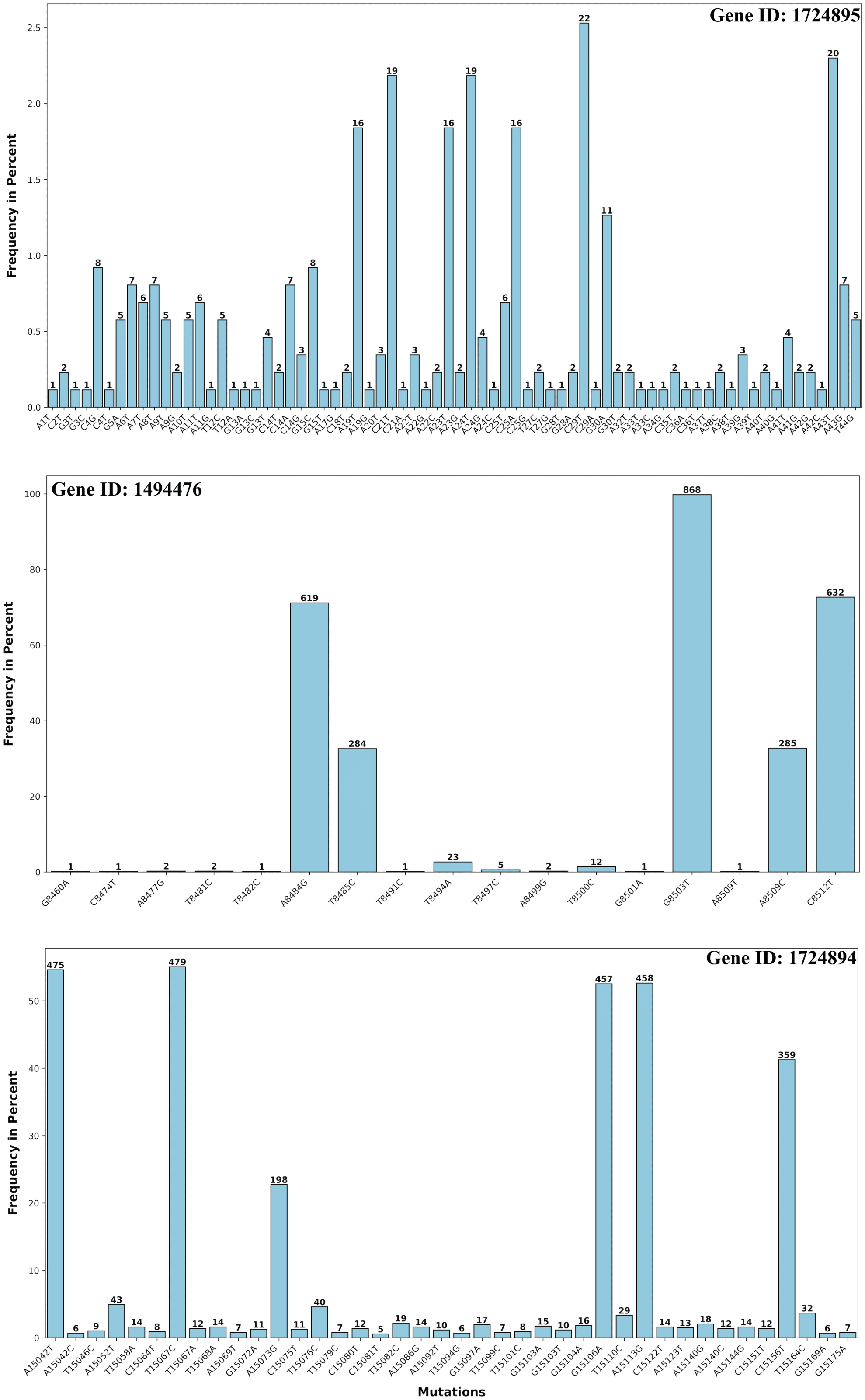

Fig 4. Nucleotide substitution frequencies of RSV non-coding segments. The analysis plot depicts the nucleotide substitution frequencies for the RSV non-coding segments (misc_RNA genes 1724895, 1494476, and 1724894),

revealing a variable range of substitution frequencies across different genomic sites. The plot highlights the heterogeneity in mutation patterns within these non-coding regions, providing insights into their evolutionary dynamics and potential functional roles.

https://doi.org/10.1371/journal.pone.0319437.g004

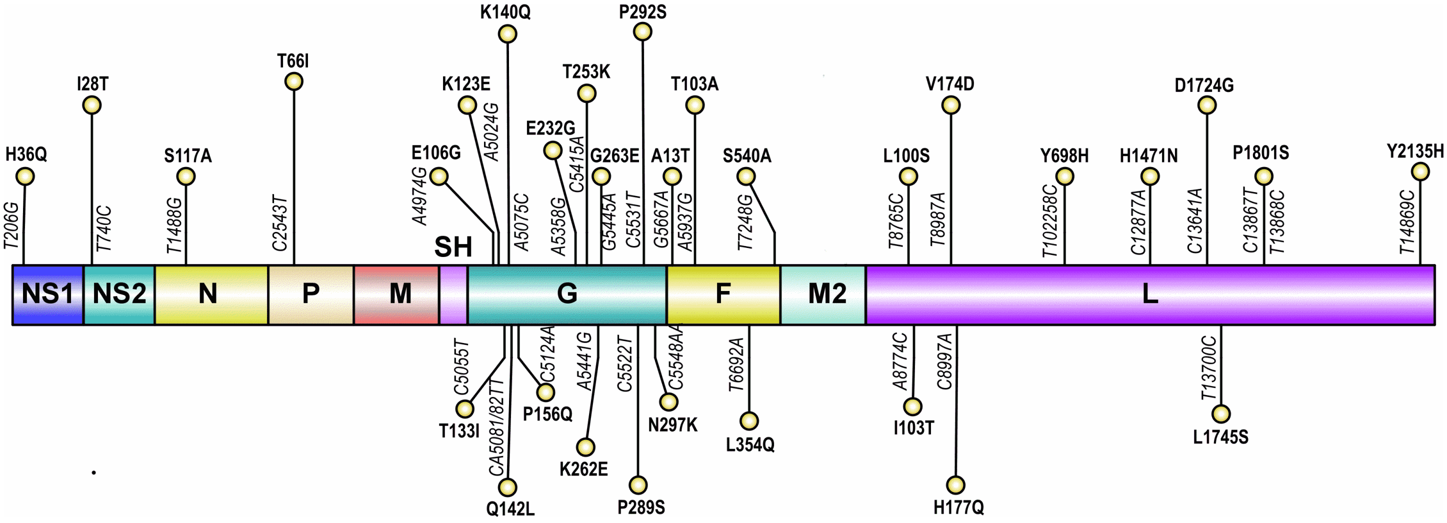

Fig 5. Schematic Presentation of hotspots mutations in RSV proteins. The nucleotide positions are numbered according to reference genome U39662.1 while the proteins are numbered as per protein.

https://doi.org/10.1371/journal.pone.0319437.g005

Substitution hotspots in RSV type A genomes

Next, we were interested in identifying the hot-spot substitution sites in RSV CDSs. Similar to the analysis [18] for SARS-CoV-2, we defined a criterion for hotspot regions. A site with a substitution frequency over 200 will be considered a hot-spot site, or a site that offered substitution to more than 200 RSV strains (42% in our case) will be considered a potential substitution hot-spot. Second, the respective site must allow non-synonymous substitution, and the observed amino acid should record a change in amino acid properties. A total of 367 substitution sites were found with > 200 substitution frequency where 290, and 77 were synonymous and non-synonymous respectively. Among 77 non-synonymous substitution sites, 31 were those affecting amino acid properties (either changing polarity or charge difference) considered hot-spot substitution sites. These 31 hotspots were found distributed across F protein (4), G protein (13), L protein (10), N1, N2, P, and N (1 each) (Fig 5). Among them, one of these hotspots that resulted in CAA to TTA (Q142L), and CAA to TCA (Q142S) in G protein offered double substitutions.

Discussion

Respiratory syncytial virus (RSV) is a communal respiratory virus that can cause mild to severe medical conditions, predominantly in young children, elderly adults, and people with certain chronic medical conditions. In some cases, it can lead to more serious illnesses such as bronchiolitis or pneumonia. Here, we detected a probable transmission pattern based on genetic global variability observed in RSV genome types A that in turn may also explain the classification of different strains circulating worldwide. The RSV evolutionary trajectories and geographical distribution are highlighted through network community clustering across different countries. Studying viral genetic variability is of utmost importance as it informs on the emergence of new escape variants and strains also deduce circulation patterns [22]. The results obtained in the current analysis are in line with the previous studies, suggesting RSV heterogeneity at both regional and temporal levels [23]. The tight clustering of the USA sequences in 2014 and 2017, which is diverged into Brazilian cluster in 2021, observed in the current analysis, depicts that viral strains may persist and evolve locally for a period before spreading to other regions. These findings are in agreement with previous reports conducted on global epidemiology of antigenically distinct viral strains of RSV, indicated a seasonally favored outbreak, dominating different regions at different times [24,25]. There are reports claiming simultaneous circulation of RSV in different communities [26]. However, in a particular region, strains of one group often dominate for one or more continuous seasons. For instance, in the year 1988 to 1990, several European countries were faced RSV epidemics caused by GA1 lineage [25]. Subsequently, in the following years RSV GA1 strains were detected less frequently in these countries. It also suggests that genetically diverse RSV strains circulate simultaneously in a locality [13]. Thus, certain RSV strains may have wide geographic dissemination and that the observed variability is predominantly temporal rather than geographic.

In viral genomes, the presence of synonymous and non-synonymous mutations with varying frequencies are primarily based on evolutionary pressure. It has been reported that in many viruses, the accumulation of non-synonymous mutations is higher compared to synonymous and is partly due to its role in immune evasion [27]. Similar results were observed in the current study where non-synonymous mutations were found partly higher (21.7%) than synonymous mutations (15.1%). Also observed mutation patterns of SARS-CoV-2 depicted higher non-synonymous mutations and were attributed to evasions schemes of host immune system [27–29].

The role of misc_RNA segments in RSV (Respiratory Syncytial Virus) genomes primarily involves the regulation of viral replication and interaction with host cells. These non-coding RNA segments play significant roles in modulating host-virus interactions and potentially influencing the virus's ability to establish infections. Various host-derived non-coding RNAs, such as miRNAs, lncRNAs, and tRNA-derived RNA fragments, are implicated in RSV infections by regulating gene and protein expression, impacting the disease mechanisms of RSV and potentially serving as biomarkers for diagnosis and targets for antiviral therapies [30,31]. Particularly, in the context of plant RNA viruses like Rice stripe virus (RSV), variations in the 3'-terminal regions of the viral genome, influenced by host alternations, play a role in viral adaptation and replication. Such variations can affect the virus's ability replicating in different hosts, such as plants and insect vectors [32].

RSV is an enveloped virus with a linear, single-stranded, negative-sense RNA genome, belonging to the *Paramyxoviridae* genus, and *Pneumoviridae* family. It retains two antigenic groups of strains, A and B, and multiple genotypes within the two groups. Structurally, RSV consists of ten genes, encoding 11 proteins. Of these eight are structural including the glycoprotein G, the fusion protein F and the hydrophobic SH protein. On the inner side of the envelope, a non-glycosylated matrix protein M is present. There are also four nucleocapsid proteins that include the nucleoprotein N, the phosphoprotein P, the transcription factor M2-1, and the large subunit of polymerase L. [33,34]. The proteins G, L and F are amongst the key proteins involved in the viral entry into the host, virus replication and immune system evasion [35].

In particular, protein G and F proteins are considered important because both can induce neutralizing antibodies, and are heavily glycosylated, which has been shown to affect with antibody recognition [36–38]. Structurally, the G protein comprises three domains: a cytoplasmic domain (1–37 amino acids), a transmembrane domain (38–66 amino acids), and an ectodomain region (67–312) [39,40]. Interestingly, all the hotspot positions we identified belong to the ectodomain of the G protein. There are individual reports in G, F, and L proteins [41–46] and we believe that this report will assist researchers to directly pick the hotspot regions for biochemical testing.

The dN/dS ratio has been studied for RSV to assess selective pressure on genes. A study on the genetic variability of the G protein gene among RSV isolates from India found that the dN/dS ratio was higher between the GA2 and GA5 genotypes (1.78), indicating greater selective pressure, compared to within the genotypes (0.69) [47]. In our case, higher nucleotide diversity in the RSV G glycoprotein gene compared to other genes was observed.

Overall, these analyses provide a complete picture of the RSV genomes, their mutations, transmission probability across the globe, and codon selective pressure analysis.

Conclusion

In conclusion, this study provides a comprehensive analysis of RSV evolution and transmission dynamics through a multi-faceted approach. By employing transmission network analysis using an effective parsimony method, we identified key transmission clusters and patterns, shedding light on the spread of RSV across populations. Additionally, the identification of hotspot regions within the genome highlighted areas of heightened variability and potential functional significance. Evolutionary analyses, including nucleotide diversity, Shannon entropy, Tajima's D, and dN/dS ratios, revealed distinct selective pressures acting on RSV genes, with the G glycoprotein gene emerging as a major driver of viral evolution. Furthermore, the inclusion of misc_RNA genes in our analyses provided novel insights into the role of non-coding regions in RSV diversity. Collectively, these findings enhance our understanding of RSV epidemiology, evolution, and adaptation, offering valuable insights for future surveillance, therapeutic development, and vaccine design.

Supporting information

S1 Table. Complete details of the substitution in each coding gene of the RSV type A. While studying this data, please take care of the INDEL event. After the INDEL event normally codon numbers of changes.
(XLSX)

S2 Table. Codon-wise selective selection details for all the RSV genes.
(XLSX)

S1 Fig. Visualization of codon under positive selection of all the RSV genes.
(TIFF)

Author contributions

Conceptualization: Ashfaq Ahmad, Shumaila Noreen.

Data curation: Sidra Majaz, Faisal Nouroz, Yingqiu Xie, Atta Ur Rehman.

Formal analysis: Ashfaq Ahmad, Sidra Majaz, Aamir Saeed, Hamid Ur Rahman, Yingqiu Xie.

Investigation: Bilal Khan, Atta Ur Rehman.

Methodology: Ashfaq Ahmad, Sidra Majaz.

Project administration: Ashfaq Ahmad, Muhammad Abbas.

Resources: Atta Ur Rehman.

Software: Hamid Ur Rahman.

Supervision: Ashfaq Ahmad.

Validation: Sidra Majaz, Muhammad Abbas, Bilal Khan, Hamid Ur Rahman, Abdur Rashid.

Visualization: Aamir Saeed, Abdur Rashid.

Writing – original draft: Sidra Majaz, Aamir Saeed, Yingqiu Xie, Atta Ur Rehman.

Writing – review & editing: Ashfaq Ahmad, Shumaila Noreen, Muhammad Abbas, Bilal Khan, Faisal Nouroz, Abdur Rashid, Atta Ur Rehman.

References

1. Arvin AM, Fink K, Schmid MA, Cathcart A, Spreafico R, Havenar-Daughton C, et al. A perspective on potential antibody-dependent enhancement of SARS-CoV-2. Nature. 2020;584(7821):353–63. https://doi.org/10.1038/s41586-020-2538-8 PMID: 32659783
2. Bohmwald K, Gálvez NMS, Canedo-Marroquín G, Pizarro-Ortega MS, Andrade-Parra C, Gómez-Santander F, et al. Contribution of cytokines to tissue damage during human respiratory syncytial virus infection. Front Immunol. 2019;10:452. https://doi.org/10.3389/fimmu.2019.00452 PMID: 30936869
3. Noor F, Saleem MH, Javed MR, Chen J-T, Ashfaq UA, Okla MK, et al. Comprehensive computational analysis reveals H5N1 influenza virus-encoded miRNAs and host-specific targets associated with antiviral immune responses and protein binding. PLoS One. 2022;17(5):e0263901. https://doi.org/10.1371/journal.pone.0263901 PMID: 35533150
4. Kant K, Lal UR, Ghosh M. Computational breakthrough of natural lead hits from the genus of Arisaema against human respiratory syncytial virus. Pharmacogn Mag. 2018;13(Suppl 4):S780–5. https://doi.org/10.4103/pm.pm_459_16 PMID: 29491633
5. Borchers AT, Chang C, Gershwin ME, Gershwin LJ. Respiratory syncytial virus--a comprehensive review. Clin Rev Allergy Immunol. 2013;45(3):331–79. https://doi.org/10.1007/s12016-013-8368-9 PMID: 23575961
6. Feng S, Hong D, Wang B, Zheng X, Miao K, Wang L, et al. Discovery of imidazopyridine derivatives as highly potent respiratory syncytial virus fusion inhibitors. ACS Med Chem Lett. 2015;6(3):359–62. https://doi.org/10.1021/acsmedchemlett.5b00008 PMID: 25941547
7. Afonso CL, Amarasinghe GK, Bányai K, Bào Y, Basler CF, Bavari S, et al. Taxonomy of the order Mononegavirales: update 2016. Arch Virol. 2016;161(8):2351–60. https://doi.org/10.1007/s00705-016-2880-1 PMID: 27216929
8. Carvajal JJ, Avellaneda AM, Salazar-Ardiles C, Maya JE, Kalergis AM, Lay MK. Host components contributing to respiratory syncytial virus pathogenesis. Front Immunol. 2019;10:2152. https://doi.org/10.3389/fimmu.2019.02152 PMID: 31572372
9. Collins PL, Fearns R, Graham BS. Respiratory syncytial virus: virology, reverse genetics, and pathogenesis of disease. Curr Top Microbiol Immunol. 2013;372:3–38. https://doi.org/10.1007/978-3-642-38919-1_1 PMID: 24362682
10. Thongpan I, Mauleekoonphairoj J, Vichiwattana P, Korkong S, Wasitthankasem R, Vongpunsawad S, et al. Respiratory syncytial virus genotypes NA1, ON1, and BA9 are prevalent in Thailand, 2012-2015. PeerJ. 2017;5:e3970. https://doi.org/10.7717/peerj.3970 PMID: 29085762
11. Mufson MA, Orvell C, Rafnar B, Norrby E. Two distinct subtypes of human respiratory syncytial virus. J Gen Virol. 1985;66(Pt 10):2111–24. https://doi.org/10.1099/0022-1317-66-10-2111 PMID: 2413163
12. 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(1):20097. https://doi.org/10.1038/s41598-019-56552-2 PMID: 31882808
13. Yu J-M, Fu Y-H, Peng X-L, Zheng Y-P, He J-S. Genetic diversity and molecular evolution of human respiratory syncytial virus A and B. Sci Rep. 2021;11(1):12941. https://doi.org/10.1038/s41598-021-92435-1 PMID: 34155268
14. Shishir TA, Saha O, Rajia S, Mondol SM, Masum MHU, Rahaman MM, et al. Genome-wide study of globally distributed respiratory syncytial virus (RSV) strains implicates diversification utilizing phylodynamics and mutational analysis. Sci Rep. 2023;13(1):13531. https://doi.org/10.1038/s41598-023-40760-y PMID: 37598270

15. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. Nucleic Acids Res. 2015;43(Database issue):D571-7. https://doi.org/10.1093/nar/gku1207 PMID: 25428358

16. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

17. Ling Y, Cao R, Qian J, Li J, Zhou H, Yuan L, et al. An interactive viral genome evolution network analysis system enabling rapid large-scale molecular tracing of SARS-CoV-2. Sci Bull (Beijing). 2022;67(7):665–9. https://doi.org/10.1016/j.scib.2022.01.001 PMID: 35036033

18. Zhou Z-J, Qiu Y, Pu Y, Huang X, Ge X-Y. BioAider: An efficient tool for viral genome analysis and its application in tracing SARS-CoV-2 transmission. Sustain Cities Soc. 2020;63:102466. https://doi.org/10.1016/j.scs.2020.102466 PMID: 32904401

19. Xie Y, Li H, Luo X, Li H, Gao Q, Zhang L, et al. IBS 2.0: an upgraded illustrator for the visualization of biological sequences. Nucleic Acids Res. 2022;50(W1):W420–6. https://doi.org/10.1093/nar/gkac373 PMID: 35580044

20. Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, et al. HyPhy 2.5-A customizable platform for evolutionary hypothesis testing using phylogenies. Mol Biol Evol. 2020;37(1):295–9. https://doi.org/10.1093/molbev/msz197 PMID: 31504749

21. Domingo E, Holland JJ. RNA virus mutations and fitness for survival. Annu Rev Microbiol. 1997;51:151–78. https://doi.org/10.1146/annurev.micro.51.1.151 PMID: 9343347

22. Tramuto F, Maida CM, Randazzo G, Guzzetta V, Santino A, Li Muli R, et al. Whole-genome sequencing and genetic diversity of human respiratory syncytial virus in patients with influenza-like illness in Sicily (Italy) from 2017 to 2023. Viruses. 2024;16(6):851. https://doi.org/10.3390/v16060851 PMID: 38932144

23. Piñana M, González-Sánchez A, Andrés C, Vila J, Creus-Costa A, Prats-Méndez I, et al. Genomic evolution of human respiratory syncytial virus during a decade (2013-2023): bridging the path to monoclonal antibody surveillance. J Infect. 2024;88(5):106153. https://doi.org/10.1016/j.jinf.2024.106153 PMID: 38588960

24. Langedijk AC, Harding ER, Konya B, Vrancken B, Lebbink RJ, Evers A, et al. A systematic review on global RSV genetic data: Identification of knowledge gaps. Rev Med Virol. 2022;32(3):e2284. https://doi.org/10.1002/rmv.2284 PMID: 34543489

25. Cantú-Flores K, Rivera-Alfaro G, Muñoz-Escalante JC, Noyola DE. Global distribution of respiratory syncytial virus A and B infections: a systematic review. Pathog Glob Health. 2022;116(7):398–409. https://doi.org/10.1080/20477724.2022.2038053 PMID: 35156555

26. Rios-Guzman E, Simons LM, Dean TJ, Agnes F, Pawlowski A, Alisoltanidehkordi A, et al. Deviations in RSV epidemiological patterns and population structures in the United States following the COVID-19 pandemic. Nat Commun. 2024;15(1):3374. https://doi.org/10.1038/s41467-024-47757-9 PMID: 38643200

27. Sohpal VK. Comparative study: nonsynonymous and synonymous substitution of SARS-CoV-2, SARS-CoV, and MERS-CoV genome. Genomics Inform. 2021;19(2):e15. https://doi.org/10.5808/gi.20058 PMID: 34261300

28. Das JK, Roy S. A study on non-synonymous mutational patterns in structural proteins of SARS-CoV-2. Genome. 2021;64(7):665–78. https://doi.org/10.1139/gen-2020-0157 PMID: 33788636

29. Hoenigsperger H, Sivarajan R, Sparrer KM. Differences and similarities between innate immune evasion strategies of human coronaviruses. Curr Opin Microbiol. 2024;79:102466. https://doi.org/10.1016/j.mib.2024.102466 PMID: 38555743

30. Eilam-Frenkel B, Naaman H, Brkic G, Veksler-Lublinsky I, Rall G, Shemer-Avni Y, et al. MicroRNA 146-5p, miR-let-7c-5p, miR-221 and miR-345-5p are differentially expressed in Respiratory Syncytial Virus (RSV) persistently infected HEp-2 cells. Virus Res. 2018;251:34–9. https://doi.org/10.1016/j.virusres.2018.05.006 PMID: 29733865

31. Wu W, Choi E-J, Lee I, Lee YS, Bao X. Non-Coding RNAs and their role in respiratory syncytial virus (RSV) and human metapneumovirus (hMPV) infections. Viruses. 2020;12(3):345. https://doi.org/10.3390/v12030345 PMID: 32245206

32. Zhao W, Yu J, Jiang F, Wang W, Kang L, Cui F. Coordination between terminal variation of the viral genome and insect microRNAs regulates rice stripe virus replication in insect vectors. PLoS Pathog. 2021;17(3):e1009424. https://doi.org/10.1371/journal.ppat.1009424 PMID: 33690727

33. Ramilo O, Rodriguez-Fernandez R, Mejias A. Respiratory syncytial virus infection: old challenges and new approaches. J Infect Dis. 2023;228(1):4–7. https://doi.org/10.1093/infdis/jiad010 PMID: 36715631

34. Anderson LJ, Jadhao SJ, Paden CR, Tong S. Functional features of the respiratory syncytial virus G protein. Viruses. 2021;13(7):1214. https://doi.org/10.3390/v13071214 PMID: 34372490

35. Shah PS, Beesabathuni NS, Fishburn AT, Kenaston MW, Minami SA, Pham OH, et al. Systems biology of virus-host protein interactions: from hypothesis generation to mechanisms of replication and pathogenesis. Annu Rev Virol. 2022;9(1):397–415. https://doi.org/10.1146/annurev-virology-100520-011851 PMID: 35576593

36. Palomo C, Cane PA, Melero JA. Evaluation of the antibody specificities of human convalescent-phase sera against the attachment (G) protein of human respiratory syncytial virus: influence of strain variation and carbohydrate side chains. J Med Virol. 2000;60(4):468–74. https://doi.org/10.1002/(sici)1096-9071(200004)60:4<468::aid-jmv16>3.0.co;2-e PMID: 10686032

37. García-Beato R, Melero JA. The C-terminal third of human respiratory syncytial virus attachment (G) protein is partially resistant to protease digestion and is glycosylated in a cell-type-specific manner. J Gen Virol. 2000;81(Pt 4):919–27. https://doi.org/10.1099/0022-1317-81-4-919 PMID: 10725417

38. Jones HG, Ritschel T, Pascual G, Brakenhoff JPJ, Keogh E, Furmanova-Hollenstein P, et al. Structural basis for recognition of the central conserved region of RSV G by neutralizing human antibodies. PLoS Pathog. 2018;14(3):e1006935. https://doi.org/10.1371/journal.ppat.1006935 PMID: 29509814

39. Langedijk JP, Schaaper WM, Meloen RH, van Oirschot JT. Proposed three-dimensional model for the attachment protein G of respiratory syncytial virus. J Gen Virol. 1996;77(Pt 6):1249–57. https://doi.org/10.1099/0022-1317-77-6-1249 PMID: 8683213

40. McLellan JS, Ray WC, Peeples ME. Structure and function of respiratory syncytial virus surface glycoproteins. Curr Top Microbiol Immunol. 2013;372:83–104. https://doi.org/10.1007/978-3-642-38919-1_4 PMID: 24362685

41. García-Barreno B, Delgado T, Melero JA. Oligo(A) sequences of human respiratory syncytial virus G protein gene: assessment of their genetic stability in frameshift mutants. J Virol. 1994;68(9):5460–8. https://doi.org/10.1128/JVI.68.9.5460-5468.1994 PMID: 8057428

42. Collins PL, Murphy BR. New generation live vaccines against human respiratory syncytial virus designed by reverse genetics. Proc Am Thorac Soc. 2005;2(2):166–73. https://doi.org/10.1513/pats.200501-011AW PMID: 16113487

43. Juhasz K, Whitehead SS, Boulanger CA, Firestone CY, Collins PL, Murphy BR. The two amino acid substitutions in the L protein of cpts530/1009, a live-attenuated respiratory syncytial virus candidate vaccine, are independent temperature-sensitive and attenuation mutations. Vaccine. 1999;17(11–12):1416–24. https://doi.org/10.1016/s0264-410x(98)00381-8 PMID: 10195777

44. Whitehead SS, Firestone CY, Collins PL, Murphy BR. A single nucleotide substitution in the transcription start signal of the M2 gene of respiratory syncytial virus vaccine candidate cpts248/404 is the major determinant of the temperature-sensitive and attenuation phenotypes. Virology. 1998;247(2):232–9. https://doi.org/10.1006/viro.1998.9248 PMID: 9705916

45. Whitehead SS, Firestone CY, Karron RA, Crowe JE Jr, Elkins WR, Collins PL, et al. Addition of a missense mutation present in the L gene of respiratory syncytial virus (RSV) cpts530/1030 to RSV vaccine candidate cpts248/404 increases its attenuation and temperature sensitivity. J Virol. 1999;73(2):871–7. https://doi.org/10.1128/JVI.73.2.871-877.1999 PMID: 9882287

46. Whitehead SS, Juhasz K, Firestone CY, Collins PL, Murphy BR. Recombinant respiratory syncytial virus (RSV) bearing a set of mutations from cold-passaged RSV is attenuated in chimpanzees. J Virol. 1998;72(5):4467–71. https://doi.org/10.1128/JVI.72.5.4467-4471.1998 PMID: 9557743

47. Parveen S, Sullender WM, Fowler K, Lefkowitz EJ, Kapoor SK, Broor S. Genetic variability in the G protein gene of group A and B respiratory syncytial viruses from India. J Clin Microbiol. 2006;44(9):3055–64. https://doi.org/10.1128/JCM.00187-06 PMID: 16954227