Diversity and Evolution of Computationally Predicted T Cell Epitopes against Human Respiratory Syncytial Virus

Short Title: Sequence-based characterization of RSV T-cell immune landscape

Jiani Chen [1][2][3][4], Swan Tan [1][3][4][5], Vasanthi Avadhanula [6], Leonard Moise [3][7], Pedro A Piedra [6], Anne S De Groot [3][7], Justin Bahl [1][2][3][4][5][8] *

[1] Center for Ecology of Infectious Diseases, University of Georgia, Athens, GA, USA

[2] Institute of Bioinformatics, University of Georgia, Athens, GA, USA

[3] Center for Vaccines and Immunology, University of Georgia, Athens, GA, USA

[4] Center for Influenza Disease and Emergence Response, University of Georgia, Athens, GA, USA

[5] Department of Infectious Diseases, University of Georgia, Athens, GA, USA

[6] Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, USA

[7] EpiVax Inc., Providence, RI, USA

[8] Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA, USA

* Corresponding author email: justin.bahl@uga.edu

Abstract

Human respiratory syncytial virus (RSV) is a major cause of lower respiratory infection. Despite more than 60 years of research, there is no licensed vaccine. While B cell response is a major focus for vaccine design, the T cell epitope profile of RSV is also important for vaccine development. Here, we computationally predicted putative T cell epitopes in the Fusion protein (F) and Glycoprotein (G) of RSV wild circulating strains by predicting Major Histocompatibility Complex (MHC) class I and class II binding affinity. We limited our inferences to conserved epitopes in both F and G proteins that have been experimentally validated. We applied multidimensional scaling (MDS) to construct T cell epitope landscapes to investigate the diversity and evolution of T cell profiles across different RSV strains. We find the RSV strains are clustered into three RSV-A groups and two RSV-B groups on this T epitope landscape. These clusters represent divergent RSV strains with potentially different immunogenic profiles. In addition, our results show a greater proportion of F protein T cell epitope content conservation among recent epidemic strains, whereas the G protein T cell epitope content was decreased. Importantly, our results suggest that RSV-A and RSV-B have different patterns of epitope drift and replacement and that RSV-B vaccines may need more frequent updates. Our study provides a novel framework to study RSV T cell epitope evolution. Understanding the patterns of T cell epitope conservation and change may be valuable for vaccine design and assessment.

Author Summary

Lower respiratory infections caused by human respiratory syncytial virus (RSV) is a global health challenge. B cell epitope immune response has been the major focus of RSV vaccine and therapeutic development. However, T cell epitope induced immunity plays an important role in the resolution of RSV infection. While RSV genetic diversity has been widely reported, few studies focus on RSV T epitope diversity, which can influence vaccine effectiveness. Here, we use computationally predicted T cell epitope profiles of circulating strains to characterize the diversity and evolution of the T cell epitope of RSV A and B. We systematically evaluate the T epitope profile of RSV F and G proteins. We provide a T cell epitope landscape visualization that shows co-circulation of three RSV-A groups and two RSV-B groups, suggesting potentially distinct T cell immunity. Furthermore, our study shows different levels of F and G protein T cell epitope content conservation, which may be important to correlate with duration of vaccine protection. This study provides a novel framework to study RSV T cell epitope evolution, infer RSV T cell immunity at population levels and monitor RSV vaccine effectiveness.

Introduction

Human respiratory syncytial virus (RSV) is a negative-strand RNA virus that is classified in the *Orthopneumovirus* genus of the family *Pneumoviridae*. It is a major cause of lower respiratory disease in young infants, immunocompromised individuals, and elderly people, resulting in annual epidemics worldwide [1]. The single-stranded RNA genome of RSV is approximate 15.2 kb and encodes 11 viral proteins [2]. The Fusion (F) and Glycoprotein (G) proteins are the two major surface proteins [3]. F protein is generally thought to be conserved and therefore it is the focus of most current RSV vaccine designs. Although G protein is highly variable, its contribution to disease pathogenesis and its role in the biology of infection suggest it can also be an effective RSV vaccine antigen [4]. Despite the significant burden of RSV infection worldwide, there is no licensed vaccine. The only approved intervention is passive immuno-prophylaxis with palivizumab, which is achieved by administering the monoclonal antibody (mAb) to a highly restricted group of infants under the age of 24 months and treatment must be repeated monthly during the RSV season due to the relatively short half-life of the antibody[5], [6]. Due to the high cost of monoclonal antibody treatments, this intervention is limited to high-risk infants and is generally unavailable in developing countries. An RSV vaccine is an urgent global healthcare priority, and it is likely that different strategies are needed for the various high-risk groups.

A number of research teams have worked on the development of RSV vaccine since its isolation and characterization in 1956 [7], [8]. However, vaccination with the formalin-inactivated, alum precipitated RSV (FI-RSV) vaccine in RSV-naïve infants and young children, led to the development of vaccine enhanced disease (VED) that hampered vaccine development for decades to follow [9]. Many studies have been conducted to explain this undesirable outcome. It is likely that formalin fixation led to a vaccine that mostly presented the post-fusion conformation of RSV F protein, leading to an excess of non-neutralizing antibodies and immune complex formation [10] [11] [12]. Other studies indicated that an impaired T cell response with Th2 skewing [13], [14], as well as complement deposition in the lungs, contributed to enhanced neutrophil recruitment [12]. Due to the recent breakthrough to structural constrain the F protein in the pre-fusion conformation and the development of RSV rodent models, there has been a surge in the number of RSV vaccine candidates undergoing clinical evaluation.

While most current RSV vaccination strategies focus on a B-cell-induced neutralization immune response, T cell immunity also plays a major role in the resolution of virus infection and is essential for RSV vaccine development [15], [16]. Once RSV infection of the lower airways is established, CD8 T cells play an important part in viral clearance and CD4 helper T cells can orchestrate cellular immune responses and stimulate B cells to produce antibodies. However, Th2-biased responses have been associated with animal models of RSV VED, and measurement of Th1 and Th2 responses are considered important to predict the safety of vaccine candidates [12]. Therefore, induction of a balanced cell-mediated immune response through vaccination would promote RSV clearance, but caution must be taken to avoid the potential for immunopathology. Taken together, a closer examination of T cell immunity and the virus sequences that induce T cell responses are needed for RSV vaccine development.

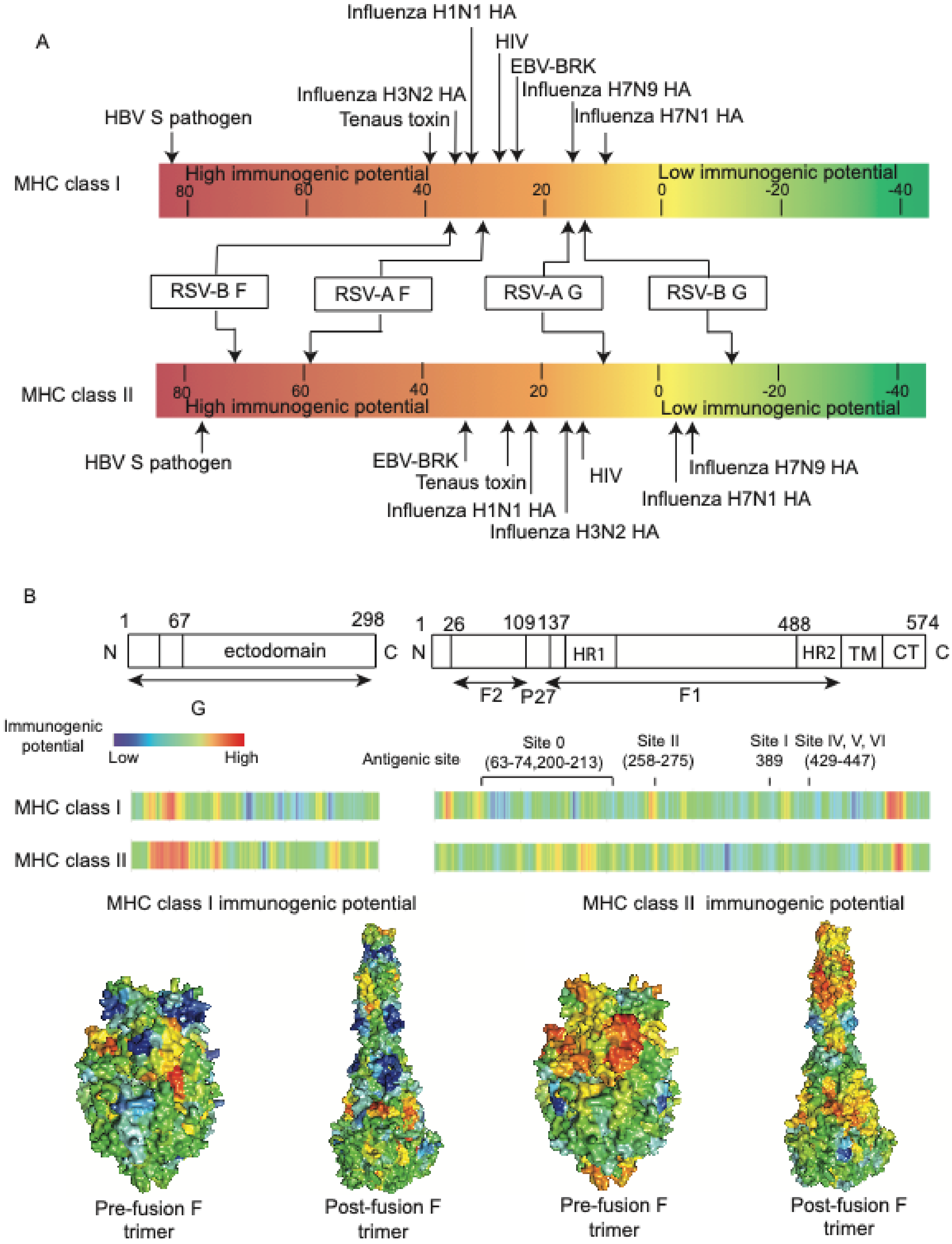
Human respiratory syncytial virus has a complex circulation pattern in the human population. Within two antigenic groups, RSV-A and RSV-B, different genotypes can co-circulate within the same community, while novel RSV genotypes with high genomic diversity may arise and potentially replace the previously dominant genotypes [17]. In recent years, several unique genetic modifications in RSV have been identified, including a 72-nucleotide (nt) duplication (ON genotype) in RSV-A G gene and another with a 60-nt duplication (BA genotype) in RSV-B at a similar region [18]. The observed RSV genetic diversity has raised a question about whether it is necessary for an RSV vaccine to include several different strains to be effective. Most current RSV vaccine developments are based on an RSV A2 laboratory strain, which is a chimeric strain that belongs to subtype A [19]. While these treatments hold promise, there is the possibility of viral strains developing escape mutations. For example, palivizumab-resistant strains have been isolated from both RSV rodent models and human [20][16]. Monoclonal antibody tests have demonstrated additional antigenic variability within RSV-A /B antigenic groups and suggest that it may play a role in the ability of RSV to escape immune response and established infections [21]. In addition, amino-acid variation at the T cell epitope level and the emergence of novel T cell epitopes have been reported [22], but further studies are needed to illustrate the effect of these variations on T cell recognition. T cell epitopes are sometimes cross-reactive, which is defined as the recognition of two or more epitope peptide-MHC complexes by the same T cell receptor and these cross-reactive epitopes are predicted to be cross-conserved [23]. Prediction of cross-conservation is important for vaccine design because it would be useful to predict protection against a pathogen with different lineages and identify escape variants. Hence, characterizing T cell epitope profiles across different strains can be crucial for RSV vaccine development.

In this study, we utilize immunoinformatic approaches that are implemented in the iVAX toolkit [24] to predict T cell epitopes in RSV across different strains with a focus on the two major surface proteins F and G. With the analysis of a comprehensive dataset, we evaluate the lineage-specific T cell epitope profile of RSV. We also create sequence-based T cell epitope landscapes based on epitope content comparison across different strains and further correlate RSV T cell immunity change with virus evolution. The proportion of cross conserved T cell epitope content between vaccine candidate strains that developed earlier and RSV circulating strains with different isolated years and locations were also calculated. These analyses may aid in understanding RSV T cell immunity across different strains and contribute to current vaccine design efforts.

Results:

Distribution of T cell epitopes in RSV surface proteins

We evaluated the T cell immunogenic potential across RSV surface proteins by scanning 9 residue regions to predict the binding probability to MHC class I and class II molecules (Figure 1). The epitope density of RSV surface proteins was evaluated using a normalized epitope density score. F protein has an epitope density score greater than +20 for both the class I and class II immunogenicity scale analysis, indicating significant immunogenic potential [24]. This contrasts with lower G protein class I and class II epitope density protein scores for both subtypes. The class I epitope density score of G protein was greater than +10 in both subtypes but the class II density was lower than random expectation in the analysis of RSV-B (Figure 1A). This result suggests that RSV surface proteins are likely to have the potential to stimulate T cells that are required for protective immunity. We then investigated the distribution of T cell immunogenicity across the proteins and found that there are regions with relatively high T cell immunogenic potential (Figure 1B). The distribution of T cell immunogenicity of F protein was mapped onto its protein structure and overlap between protein sequence regions with high T cell immunity potential and the antibody neutralizing targets was observed at antigenic site Φ and site II.


A
Influenza H1N1 HA
HIV
EBV-BRK
Influenza H7N9 HA
Influenza H7N1 HA
Influenza H3N2 HA
Tenaus toxin
HBV S pathogen
MHC class I
High immunogenic potential
Low immunogenic potential
80
60
40
20
0
-20
-40
RSV-B F
RSV-A F
RSV-A G
RSV-B G
MHC class II
HBV S pathogen
EBV-BRK
Tenaus toxin
Influenza H1N1 HA
Influenza H3N2 HA
HIV
Influenza H7N9 HA
Influenza H7N1 HA
B
1
67
298
N
ectodomain
C
G
1
26
109
137
488
574
HR1
HR2
TM
CT
F2
P27
F1
Immunogenic potential
Low
High
Antigenic site
Site 0 (63-74,200-213)
Site II (258-275)
Site I 389
Site IV, V, VI (429-447)
MHC class I
MHC class II
MHC class I immunogenic potential
MHC class II immunogenic potential
Pre-fusion F trimer
Post-fusion F trimer
Pre-fusion F trimer
Post-fusion F trimer

146

Figure 1: T cell immunogenic potential for RSV surface proteins based on MHC binding prediction. (A) T cell immunogenic potential of RSV major surface proteins. T cell epitope density scores for RSV major surface proteins and other pathogen proteins are labeled on a scale bar. Low-scoring proteins are known to engender little to no immunogenicity while higher-scoring proteins are known immunogens. Proteins scoring above +20 on this scale are considered to have significant immunogenic potential. (B) Distribution of RSV T cell immunogenic potential across F and G protein in RSV reference strain A2. Prefusion or post-fusion F protein surface was colored by the relative immunogenetic potential at each location. Analyses are based on the RSV-A reference sequence.

Lineage specific T cell epitope profiles

We then extended T cell epitope predictions from RSV representative strains to multiple wild-circulating strains. The distribution and diversity of T cell epitopes across different strains are illustrated in heatmaps with the corresponding time-scaled phylogenies (Supplementary Figure 1 and Supplementary Figure 2). Both F and G proteins contain epitopes that were conserved across all RSV strains in almost 100% of sampled isolates, suggesting that they could serve as high-quality T cell epitope candidates for vaccine design. In contrast, some epitopes were mutated in selected strains, and those epitopes that only occurred in certain clades within the phylogeny could be interpreted as clade-specific “fingerprints”.

The G gene duplication events in RSV, which are unique gene signatures, can either shift the position of epitopes or cause the emergence of novel epitopes. Two novel class I epitopes, (no. 31 and no. 40 in Supplementary Figure 2A), were found in RSV-A strains that contain G gene duplication. In addition, an emergent class II epitope (no. 25 in Supplementary Figure 2A) was identified in RSV-A sequences that contain G gene duplication, which was a shift from an epitope (no. 24) that has been observed in other strains. From RSV-B strains that contain the G gene duplication event, we also observed multiple lineage specific class I T cell epitopes, which are caused by a 2-aa deletion (aa157 and aa158) in these strains instead of directly due to the 60-nt duplication event. RSV-B G proteins that have the duplication event contain multiple novel epitopes (no. 22, 23, 26, 28, 30, 37) but do not contain several epitopes (no. 24, 25, 27, 29, 31, 38) that are identified in other strains (Supplementary Figure 2B).

To further determine whether the T cell epitopes defined using EpiMatrix might be immunogenic, we utilized the JanusMatrix [25] algorithm to identify the T cell epitopes that are likely to be cross-conserved with human epitopes and thereby tolerated by the immune system. Based on this analysis, 6.45% of putative class I epitopes and 1.12 % of putative class II epitopes of RSV major surface proteins are cross-conserved with human proteome-derived epitopes at TCR facing residues. As these peptides have similar HLA binding preferences that are contained in human proteins (Supplementary Figure 3), they were therefore assumed not to be immunogenic. After excluding the high-JanusMatrix score epitopes identified above, we were able to identify T cell epitopes that were conserved in more than 60% of currently circulating RSV strains. We searched the IEDB epitope database to determine if these epitopes were related to experimentally validated RSV T cell epitopes or HLA ligands. The conserved RSV T cell epitope sequences that may be important for future vaccine development are shown in Table 1 and Table 2.

Table 1: Experimentally validated conserved MHC class I epitopes peptides in RSV major surface proteins [a]

Subgroup	Protein	Epitope address	Epitope sequence [b]	Binding HLAs [c]	Conservation [d]	Number of human matches [e]	Epitope id in IEDB
RSV-A & RSV-B	F	45-53	**LSALRTGWY**	A0101	99.55%(A) & 74.24%(B)	1	158982
		140-148	**FLLGVGSAI**	A0201	99.59%(A) & 97.98%(B)	0	156869
		250-258	**YMLTNSELL**	A0201, A2402	99.59%(A) & 99.33%(B)	0	156979
		272-280	**KLMSSNVQI**	A0201	66.64%(A) & 96.08%(B)	3	156902
		273-281	**LMSSNVQIV**	A0201	66.56%(A) & 96.08%(B)	1	156915
		449-457	**TVSVGNTLY**	A0101	99.75%(A) & 99.33%(B)	0	97017
RSV-A	F	10-18	AITTILAAV	A0201	84.69%	3	156844
		111-119	LPRFMNYTL	B0702	91.18%	0	158975
		170-178	ALLSTNKAV	A0201	99.67%	2	156847
		383-391	NIDIFNPKY	A0101	95.86%	0	159045
	G	25-33	FISSCLYKL	A0201	99.26%	0	158759
		61-69	FIASANHKV	A0201	82.08%	0	158751
RSV-B	F	525-533	IMITAIIIV	A0201	89.25%	0	156892
		540-548	SLIAIGLLL	A0201	97.65%	5	156960
	G	25-33	VISSCLYKL	A0201	90.91%	0	158759
		61-69	FIISANHKV	A0201	99.02%	0	158751

a. This table contains putative MHC class I epitopes that have already been experimentally validated in publications.

b. Epitopes sequences that are conserved in both RSV-A and RSV-B are in bold.

c. HLAs that have the top 1% binder scores in EpiMatrix for epitope sequence.

d. The conservation is evaluated by the presence of epitope peptides across all RSV-A or RSV-B sequences that are publicly available (only epitope sequences with at least 60% conservation are shown in the table).

e. Count of human epitopes found in the search database. JanusMatrix was used to search human epitopes that are predicted to bind to the same allele as the RSV epitope and share TCR facing contacts with the RSV epitope.

Table 2: Experimentally validated conserved MHC class II epitopes peptides in RSV major surface proteins [a]

Subtype	Protein	Epitope address	Epitope sequence [b]	Conservation [c]	Number of human matches [d]	Epitope id in IEDB
RSV-A	F	29 - 44	TEEF**YQSTCSAVS**KGY	98.53%	3	956680
		50 - 70	TGW**YTSVITIELSNIK**ENKCN	97.75%	1	153700
		167 - 192	IKSALLSTNKAVVSLSNGVSVLTSKV	93.14%	4	545502
		218 - 234	ETVIEFQQKNNRLLEIT	98.86%	3	1087566
		247 - 268	VSTYMLTNSELLSLINDMPITN	98.98%	8	99471
		288 - 310	IMSIIKEEVLAYVVQLPLYGVID	98.57%	5	99334
		399 - 418	KTDVSSSV**ITSLGAIVS**CYG	99.14%	0	545603
		453 - 470	GNTLYYVNKQEGKSLYVK	98.37%	1	99691
		492 - 510	ISQVNEKI**NQSLAFIR**KSD	80.32%	1	153713
		543 - 560	AVG**LLLYCKARSTPV**TLS	79.26%	6	153641
	G	19 - 43	TLNHLLFISSCLYKLNLKSIAQITL	93.13%	8	1087567
RSV-B	F	29 - 44	TEE**FYQSTCSAVS**RGY	99.78%	3	956680
		50 - 70	TGW**YTSVITIELSNIK**ETKCN	93.95%	1	153700
		192 - 218	VLDLKNYINNQLLPIVNQQSCRISNIE	83.43%	4	153636
		247 - 268	LSTYMLTNSELLSLINDMPITN	98.54%	8	99471
		399 - 418	KTDISSSV**ITSLGAIVS**CYG	98.88%	0	545603
		453 - 470	GNTLYYVNKLEGKNLYVK	98.77%	0	99691
		492 - 510	ISQVNEKI**NQSLAFIR**RSD	97.42%	1	153713
		543 - 560	AIGL**LLYCKAKNTPV**TLS	94.96%	4	153641
	G	51 - 74	STSLIIAAIIFIISANHKVTLTTV	94.66%	8	158751

a. This table contains putative MHC class II epitopes which share the identical binding groove sequence of the RSV class II epitopes that have already been experimentally validated in publications.

b. Underlined sequences represent the nine-mer frames with the greatest potential to bind class II HLA. Epitope sequences that are in bold indicate sequences are predicted to bind class II HLA and are conserved in both RSV-A and RSV-B.

c. Conservation is evaluated by the presence of epitope peptides across all RSV-A or RSV-B sequences that are publicly available (Only epitope sequences with at least 60% conservation are shown in the table).

d. Count of human epitopes found in the search database. JanusMatrix was used to search human epitopes that are predicted to bind to the same allele as the RSV epitope and share TCR facing contacts with the RSV epitope.

Predicted RSV T cell epitope landscapes

To investigate the evolution of RSV on T cell immunity profiles, we develop a new approach to visualize the T immunity profile of multiple RSV strains on a landscape. We performed a T cell epitope content pairwise comparison between RSV strains, and T cell epitope distances between pair of RSV strains were then calculated using a T epitope distance algorithm. We then applied a multidimensional scaling (MDS) approach using these estimated pair-wise T epitope distances to map RSV strains to a landscape to characterize their T-cell immunity profile. We found both Class I and Class II T cell immunity profiles of F and G proteins of different RSV strains were clustered into groups on this T cell epitope landscapes (Supplementary Figure 4). Combining the Class I and Class II T-cell epitope biding profiles, RSV-A major surface protein isolates can be divided into three clusters and RSV-B major surface protein isolates can be divided into two clusters (Figure 2, Supplementary Figure 5). We observe that the G gene sequence isolates that contain 72-nt (RSV-A) or 60-nt (RSV-B) duplications clustered together with other sequences instead of forming isolated groups. To further investigate the T cell epitope diversity, we correlated this clustering pattern with the phylogenetic histories (Figure 2B). The phylogenetic tree topologies of the RSV-A F gene and G gene are similar. The F gene cluster 1 is paraphyletic, while cluster 2 and 3 are monophyletic. Cluster 1 is the closest to the ancestral sequence and mapping this group onto the phylogeny show that this cluster has a basal relationship with clusters 2 and 3 indicating that the phylogenetic divergence occurred prior to epitope drift. The RSV-B F and G gene genealogies are very different. In particular, the RSV-B F gene topologies is indicative of strong immune selection, similar to observed human influenza A virus or within host HIV phylogenies [26]. In contrast, the RSV-B G gene phylogeny shows the co-circulation of multiple lineages, though this could reflect the sequencing bias of G genes (Figure 2B). We then calculated the T-cell epitope immune distance of each strain from a reconstructed ancestral sequence (Figure 2C). These distances were then plotted against the year of isolation and colored according to the cluster identified in Figure 2A. RSV-A shows that multiple predicted immune phenotypes co-circulate and persist for long periods (>2 decades). Analysis of RSV-B shows a turnover of the predicted immune phenotypes with short periods of co-circulation (<5 years) for F and G protein T cell epitopes. The limited periods of co-circulation is again consistent with phenotype patterns observed for viruses under strong immune selection (e.g H3N2 influenza A virus) [27], [28]. In contrast, genetic distances from the reconstructed ancestral sequence plotted against year of isolation show patterns typical of gradual genetic drift, except in the G gene where a 72-nt and 60-nt insertion is present (Figure 2D). Taken together, these results suggest that genetic and predicted T-cell epitope immune diversity are different and may be an important factor to consider when evaluating RSV vaccine efficacy.

There are multiple methods available to predict T cell epitopes [29], which may result in different reconstructed landscapes if there is a systematic bias in the prediction method. We used the NetMHCpan method [30] to predict T cell epitopes and perform the same landscape reconstruction using MHC class I binding predictions for RSV-A F protein. Our analysis showed a consistent clustered pattern of RSV T epitope profile on the landscape regardless of T cell epitope prediction method (Supplementary Figure 6).

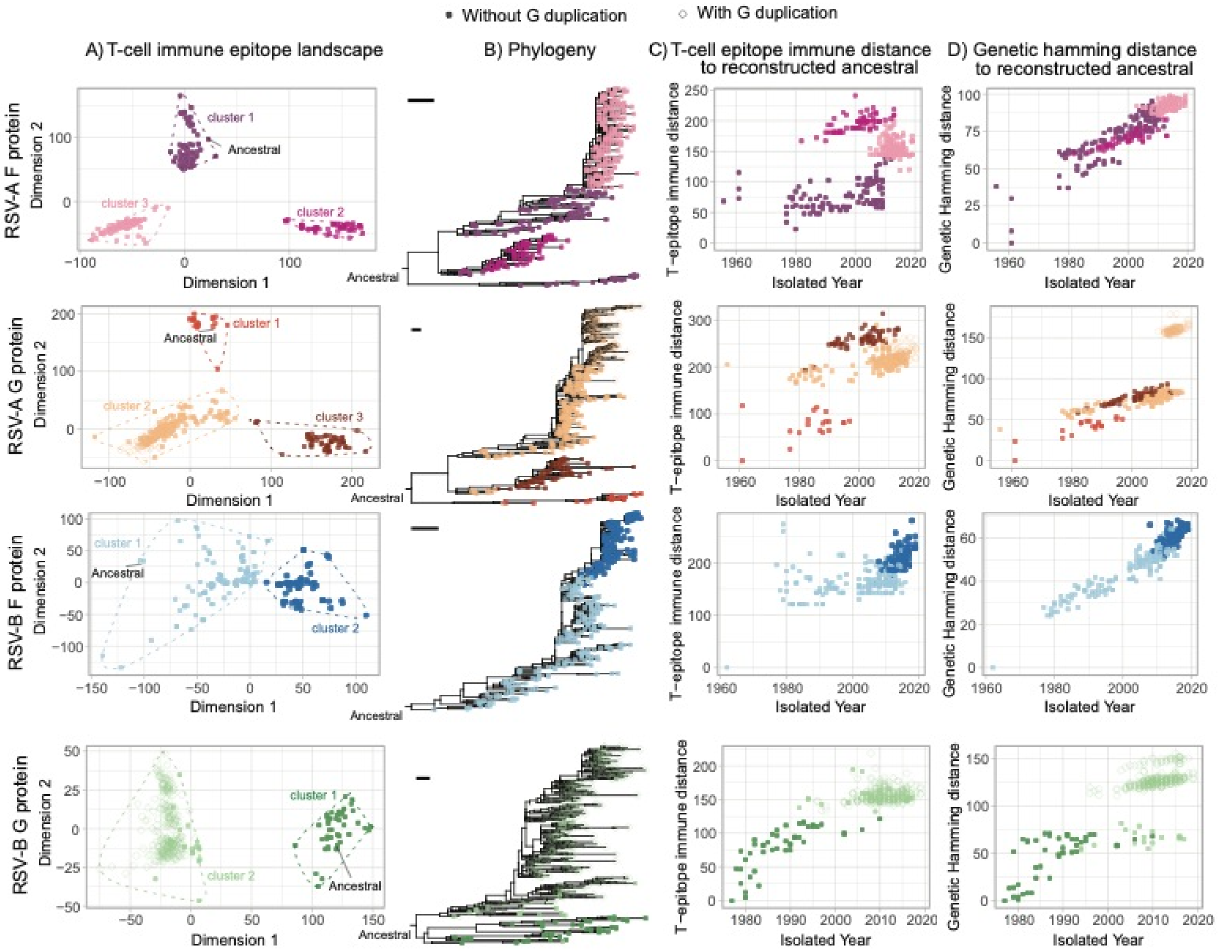


Figure 2: Predicted T cell epitope landscapes and genetic evolution of RSV surface proteins. Epitope landscapes of RSV major surface proteins are built with MHC class I and class II epitope content comparison across different strains. Filled circles indicate F protein isolates or G protein isolates without duplication. Diamonds indicate G protein isolates with gene duplication. T cell immunity clusters are determined with *k-means* method and are used to color the sequenced isolates in the following panels. The corresponding Maximum Likelihood (ML) phylogenies are reconstructed and are rooted by mid-point. Scale bars indicate 0.005 nucleotide substitution per site. T cell epitope immune distance and genetic hamming distance from the estimated TMRCA are plotted against the isolated time of each sequence.

Assessment of vaccine candidate strains with T cell epitope content

To quantitatively evaluate whether it might be necessary to include multiple RSV strains to prepare an effective vaccine, two live attenuated RSV strains that are previously considered as vaccine candidates, CP248, a recombinant virus that belongs to subtype A, and CP52, which is a recombinant RSV-B strain, were included in our analysis and compared to wild-type strains using EpiCC. We calculated the average proportion of cross-conserved T cell epitope content between the selected vaccine strains and wild-circulating strains from different isolation years and WHO regional groups (Figure 3, Supplementary Figure 7). Different proportions of cross-conserved T cell epitope content against isolates from two different subtypes, A and B, were observed in both the F and G protein analyses. In the comparison of the vaccine strains and wild strains belonging to the same subtype, the proportion of cross-conserved T cell epitope in RSV F protein is relatively stable in different groups, all are higher than 78% for RSV-A and higher than 85% for RSV-B. In contrast, changes in the proportion of cross-conserved T cell epitopes were detected among groups within the same RSV subtype, especially in different temporal groups in the G protein analysis (Figure 3). Vaccine strain CP248 appears to have a relatively higher proportion of cross conserved T cell epitopes within G protein when compared to the RSV-A strains that were isolated before 1991 (> 70%) and a relatively lower degree of conservation against recently isolated strains. A similar decrease in T cell epitope conservation with time was identified for vaccine strain CP52 among circulating RSV-B strains.

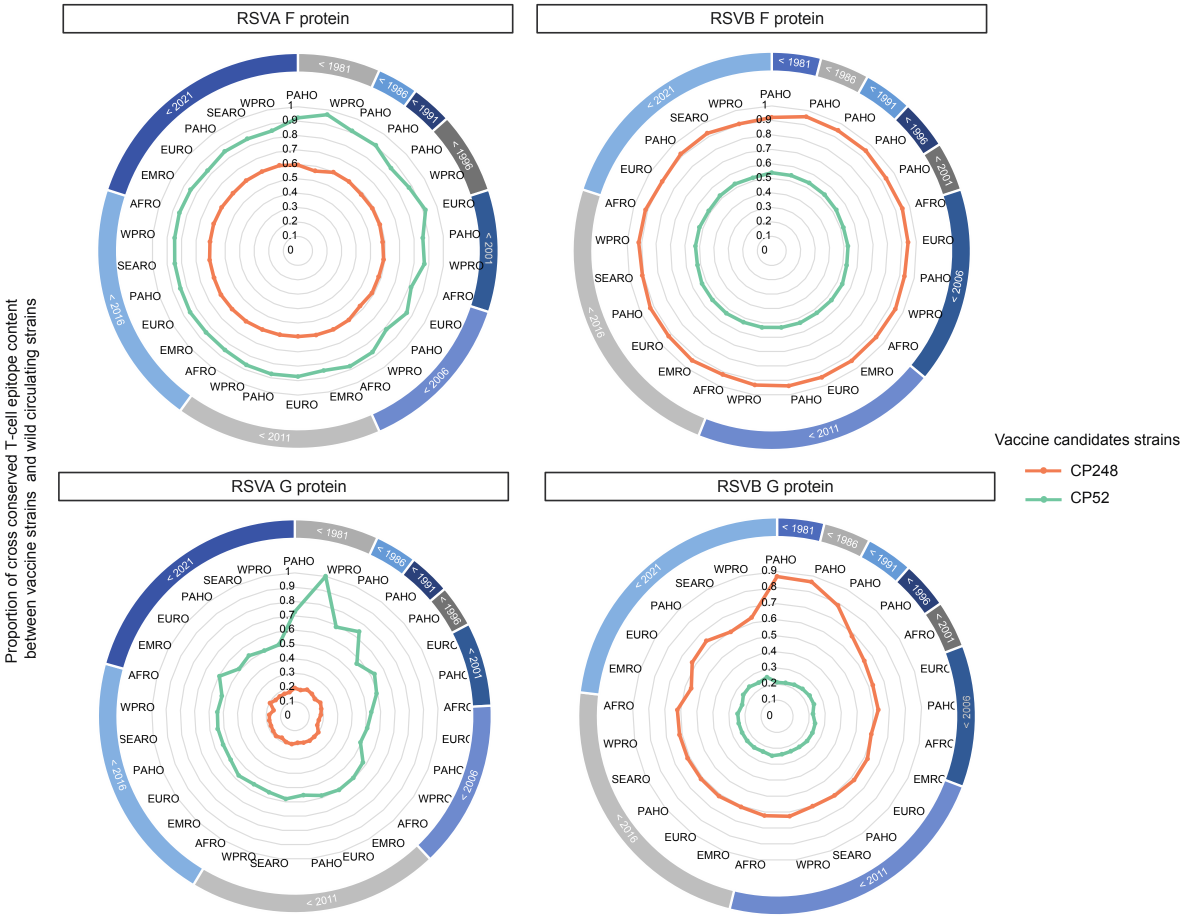


Figure 3: Evaluation of previously used RSV vaccine candidate strains with T cell epitope content of circulating strains. RSV-A and RSV-B major surface protein sequences are subsampled and then grouped by isolation year and 6 isolated WHO regions. African Region (AFRO), Region of the Americas (PAHO), South-East Asia Region (SEARO), European Region (EURO), Eastern Mediterranean Region (EMRO) and Western Pacific Region (WPRO). The proportion of cross-conserved T cell epitope content between live attenuated strains (CP248 or CP52) and wild circulating strains are displayed as radar plots.

Discussion

Although both CD4 and CD8 T cells contribute to protection against RSV-induced disease following primary infection [16][31], T cell epitopes have received limited attention in the RSV research effort. We demonstrate RSV surface proteins appear to have significant potential to drive T cell immunity using a computational approach, based on their T cell epitope density scores as determined by MHC molecular binding prediction. The relatively high putative T-cell epitope density might make F protein a good target for RSV vaccine. In addition to the analysis of T cell epitope density and distribution in RSV major surface proteins, we also demonstrated lineage-specific variations in T cell epitope content. Even though RSV F protein is believed to be well conserved and G protein is reported to be highly variable, epitope mutations are observed across different lineages within the F protein, and potential conserved T cell epitopes can still be found in the highly variable G protein, suggesting that studying the lineage-specific T cell epitopes in RSV can provide insight into the impact of immune selection on viral diversity and persistence. While experimental validation is needed, this analysis highlights the importance of understanding population-level epitope conservation as it may provide important insight into the development of T cell epitope-driven vaccines against RSV infection.

A major focus of our work is the development of a sequence-based method to map the evolution of T cell immunity across different stains. Following a previously pivotal work that used MDS method to map the evolutionary adaptation of influenza A virus-induced by CD8 T cell using the presence and absence of MHC class I epitopes [32], we constructed RSV T cell immunity landscapes using immune distances that were generated by T cell epitope cross-conservation analyses, which allows for easy visualization and intuitive understanding of the potential for T cell immunity relationships among different strains. When comparing across strains, we found that the T cell epitope content of RSV surface proteins from different strains can be clustered, as has been observed for the antigenic relationship reported in other pathogens [49][50]. Our results also demonstrate the correspondence between RSV T cell immunity clusters and their corresponding phylogeny, with sequences in the same clade generally belonging to the same T cell immune cluster. Importantly, we also observe different patterns of T-cell epitope evolution of RSV wild strains compared with their genetic evolution, which highlights the importance of characterizing T cell epitope changes in RSV.

We identified highly conserved RSV T cell epitopes in this study, some of which have already been experimentally validated and published in the IEDB database. However, we also identified several other conserved T cell epitopes that have not been previously described. These may be valuable for vaccine design, although experimental validation will be needed. Furthermore, the homology of selected RSV epitopes to human epitopes suggests that some predicted RSV T cell epitopes might be tolerated by the human immune system, or could induce a harmful cross-reactive immune response against human proteins when administered with an adjuvant [25]. Certain aspects of immunity to RSV were not addressed by this study. For example, neutralizing antibody responses are currently considered to be the most important correlate of immunity. While neutralizing antibodies would not directly be elicited by a T cell epitope-driven vaccine, helper (CD4) T cell epitopes are required to generate high affinity, high specificity antibodies. We also note that we have limited our focus on the two major RSV surface proteins in our current analysis, but other RSV proteins like N, M, or M2-2 proteins might also contribute to vaccine efficacy [36].

An effective vaccine against variable viruses should contain T cell epitopes that are highly conserved among circulating strains [37]. Vaccine efficacy can be diminished if T cell epitopes in a vaccine strain do not match when new strains of pathogens emerge. In this study, we used an immunoinformatic-based approach to estimate cross-conserved T cell epitope contents between two live attenuated vaccine candidate strains and RSV circulating wild strains. We found that there was a low proportion of cross-conserved T cell epitope content with vaccine strains that belonged to different antigenic groups, which indicates the risk of using a single-subtype strain in RSV vaccines. In addition, we observed a lower proportion of cross-conserved G-protein T cell epitope content between vaccine strains and recent circulating strains in the same antigenic group, which suggests that including T cell epitopes from different strains in the same antigenic group might also be important for RSV vaccine development. Although we did not observe a significant change in cross-conserved T cell epitope content in F protein, we cannot rule out the probability that variation of F protein in the future could render a single-strain-based vaccine less effective. Our current analysis is based on reduced datasets due to the heavy computational capacity required to perform epitope content comparison. We constructed these representative datasets by randomly subsampling the complete datasets according to geographical regions and isolated years. Our findings may reflect the T cell epitope diversity of publicly available RSV strains, however, additional RSV surveillance efforts may be required to get a full picture of the T cell epitope variability of RSV.

The lack of experimental data might cause problems in epitope identification. Computational-based T cell epitope landscapes have the potential for bias. But the observed clustered T cell evolutionary pattern in RSV surface proteins provide valuable insights into virus evolution in the aspects of T cell immunity and strain selection for vaccine design.

Overall, this study provides a focused analysis of T cell epitopes in RSV major surface proteins using computational tools. We performed a comprehensive T cell epitope prediction for RSV showing the immunological relationship of T cell epitopes in RSV surface proteins. This study demonstrates that T cell epitope evolution may differ from genetic variation and provides a framework for developing an integrated epitope-based RSV vaccine and evaluation methods that could be used to optimize vaccination strategies.

Materials and Methods

Dataset

RSV GenBank records files were retrieved from NCBI's GenBank nucleotide database using the search term "HRSVA" or "HRSVB" on June 22, 2020. F and G gene nucleotide sequences and metadata including country of isolation and collection date were extracted using customized python scripts. Genotype assignments were made with the program "LABEL", using a customized RSV module [38] [39]. Countries of isolation were grouped into 6 WHO regions: African Region, Region of the Americas, South-East Asia Region, European Region, Eastern Mediterranean Region, and Western Pacific Region [40]. The following inclusion and exclusion criteria were applied: (i) each sequence needed to have a known isolated geographic location and isolated year, (ii) each sequence had to be at least 80% of the complete gene sequence in length, (iii) identical sequences with the same isolate country were removed, and (iv) vaccine derivative and recombinant sequences were removed. Using these criteria, comprehensive datasets of RSV

F and G genes were defined (RSV-A F gene = 1010, RSV-B F gene = 894, RSV-A G gene = 1488, RSV-B G gene = 1120). Nucleotide sequences from each dataset were aligned using MAFFT.v7 [41] and were translated into amino acids using EMBOSS.v6.6.0 [42] for immunoinformatic analyses. In addition, two artificial sequences, CP248 and CP52 (cold passage live RSV strains that were previously evaluated as vaccine candidates, Accession No: U63644, AF0132551 respectively) were downloaded from the NCBI's GenBank nucleotide database [43].

Phylogenetic inference

The nucleotide sequences of RSV major surface proteins were used to reconstruct the maximum-likelihood (ML) phylogeny of RSV using RAxML.v8 with GTR+GAMMA substitution model [44]. The best-scoring ML tree was automatically generated from five runs by RAxML. Time-scaled phylogenies were further reconstructed with the best scoring ML trees using the program "Timetree" [45]. The phylogenies are visualized in the R package "ggtree" [46].

T cell epitope prediction

RSV major surface protein sequences were scored for binding potential against a globally representative panel of Human Leukocyte Antigen (HLA) class I and class II alleles using the EpiMatrix algorithm. This algorithm as well as the ClustiMer, JanusMatrix, and EpiCC algorithms discussed below are part of the iVAX toolkit developed by EpiVax, which is available for use under a license or through academic collaborations [24].

Evaluation of class I epitopes was made based on predictions for four HLA-A and two HLA-B supertype alleles: A*01:01, A*02:01, A*03:01, A*24:02, B*07:02, B*44:03. Class II epitopes were identified for nine HLA-DR supertype alleles: DRB1*01:01, DRB1*03:01, DRB1*04:01, DRB1*07:01, DRB1*08:01, DRB1*09:01, DRB1*11:01, DRB1*13:01, and DRB1*15:01.

EpiMatrix parsed 9-mer sequence frames (each one overlapping the previous one by one amino acid) from the antigen sequence and assigned a score for each nine-mer/allele pair on a normalized Z distribution. Nine-mer sequences that had Z-scores of at least 1.64 are considered to be in the top 5% of any randomly generated set of 9-mer sequences and to have a high likelihood of binding to HLA molecules and being presented to T cells. Sequences that score above 2.32 on the Z-scale (top 1%) are extremely likely to bind to a particular HLA allele and to be immunogenic. For this analysis, HLA-class I restricted 9-mer sequences that had top 1% binder scores to at least one HLA class I supertype allele were considered to be putative class I epitopes [24].

To identify putative class II epitopes, we used an algorithm called ClustiMer [24] to screen EpiMatrix scoring results for the nine class II alleles. ClustiMer identifies contiguous regions of 15–30 amino acids that have a high density of MHC class II binding potential. Epitope density within a cluster is reported as an EpiMatrix Cluster Score, where scores of 10 and above are likely to be recognized in the context of multiple class II alleles and to be high-quality class II epitopes.

We also applied analysis of human homology to this study. The JanusMatrix algorithm was used to assess the potential cross-conservation of T cell epitopes with epitopes restricted by the same HLA alleles in the human proteome [25]. Briefly, JanusMatrix scans each identified epitope and examines the shared T cell receptor (TCR) contacts with class II epitopes present in the human proteome, to compute a JanusMatrix Human Homology Score. As defined in retrospective studies, foreign class I epitopes that score greater than 2 and class II epitopes that score greater than 5 may be less immunogenic due to T cell tolerance.

Protein-level T cell immunogenic potential evaluation

RSV reference sequences (RSV-A: NC_038235, RSV-B: NC_001781) were downloaded from the NCBI RefSeq database and were used to evaluate the protein-level immunogenic potential of RSV major surface proteins. The protein-level immunogenic potential as represented by the EpiMatrix-defined T cell epitope density score was computed by summing the top 5% binder scores across HLA alleles and normalizing for a 1000-amino acid protein length. Zero on this scale is set to indicate the average number of top 5% binders that would be observed in 10,000 random protein sequences with natural amino acid frequencies. Proteins scoring above +20 have been observed to have the significant immunogenic potential [47]. Fully human proteins generally score lower than zero on the EpiMatrix immunogenicity scale.

To investigate the distribution of T cell immunogenic potential across RSV protein sequence regions, we summed up the binding scores of HLA alleles for each nine-mer frame, to get a frame-specific immunogenic potential score and standardized this score to a relative scale. The relative immunogenic potential across protein structure was represented by a color scale and the visualization of F protein structure was built with PyMOL Molecular Graphics System, Version 2.0 (Schrödinger, LLC). Protein data bank (PDB) files 5UDE [48] and 3RRR [49] were used for the pre-fusion and post-fusion forms.

Subsampling strategy

Considering the heavy computational load that would be required to evaluate all available RSV sequences and to correct the overrepresentation of recently sampled strains, the comparative analysis for T cell epitope content was conducted with datasets in which overrepresented groups were reduced. A maximum of five sequences of each isolation year from different WHO region groups were subsampled randomly from the original datasets (RSV-A F gene = 402, RSV-B F gene = 319, RSV-A G gene = 390, RSV-B G gene = 359).

T cell epitope content comparison

The Epitope Content Comparison (EpiCC) algorithm, which is implemented in iVAX was used to evaluate pairwise T cell epitope cross-conservation potential within each subsampled dataset [50]. Briefly, T cell epitope cross-conservation was evaluated by the binding likelihood of epitopes from different antigens with identical T cell receptor-facing residues (TCR*f*), which are predicted to bind to the same MHC allele. Two epitope sequences are assumed to be potentially cross-conserved if they have identical TCR*f* (position 4, 5, 6, 7, 8 for class I epitopes binding core and 2, 3, 5, 7, 8 for class II epitopes binding core) regardless of differences on their MHC-facing amino acids. To simplify the analysis, the binding of 9-mer epitopes within protein sequences are assumed to be mutually exclusive and uniform.

Therefore, T cell epitope immune distance (D) between two wild circulating strains (w_1 and w_2) can be defined as the sum of Z-scaled binding probabilities of paired epitopes that are unable to induce cross-reactivity (non-cross conserved epitopes) within two protein sequences using equations (1.1 and 1.2). d is the T cell immunity distance between a pair of epitopes, i and j are the non-cross conserved T cell epitopes from two protein sequences, a is a class I or class II allele, p is the predicted binding probability against allele a, A is a set of alleles.

$$d(i,j) = p(i)a + p(j)a\#(1.1)$$

$$D(w_1,w_2) = \sum_{i\in w_1 j\in w_2} \sum_{a\in A} d(i,j)\ \#(1.2)$$

To evaluate the T cell epitope immune distance generated by the EpiCC algorithm, we further adapt the equations (1.1 and 1.2) to re-calculate T cell epitope immune distance with customized Python scripts using MHC binding prediction results that are generated from publicly available T cell epitope prediction tool, netMHCpan EL 4.1 methods in the Immune Epitope Database (IEDB) [51]. Eigenvalues of each sequence that were calculated from the pairwise distance matrix with “RSpectra” package were used to statistically examine the correlation of the epitope distances that are computed from the two methods, and Pearson correlation test was used to test the correlation hypothesis.

The cross-conservation of vaccine strains against circulating RSV can be evaluated by the cross-conservation of the epitopes within vaccine strains (v) and wild circulating strains (w). T cell cross-conservation between two epitopes can be represented by a joint probability estimation and therefore T cell cross-conservation between two sequences can be represented by summing T cell cross-conservation of the paired T-epitopes within two protein sequences. The proportion of T cell cross-conservation between the vaccine and circulating strains (P) with a set of alleles (A) can be represented as the equations (2.1 and 2.2), where p is the predicted binding probability in EpiMatrix, i and j are the cross conserved T cell epitopes, a is a class I or class II allele.

$$S(i,j) = p(i)a * p(j)a\#(2.1)$$

$$P(v,w) = \frac{\sum_{i \in v, j \in w} \sum_{a \in A} S(i,j)}{\sum_{i \in v, j \in \text{v}} \sum_{a \in A} S(i,j)} \#(2.2)$$

Dimension reduction

The equation to calculate T cell epitope immune distance was applied iteratively to the subsampled dataset and therefore the pairwise T cell epitope immune distances are structured into an *n x n* square-distance matrix. Given that each protein is described by a relative distance to the rest of n-1 proteins, the data must be dimensionally reduced to be graphed. Classic (metric) multidimensional scaling (MDS) can be used to preserve the distances between a set of observations in a way that allows the distances to be represented in a two-dimensional space. MDS was performed as previously described by Gower [52]. The MDS method first constructs an n-dimensional Euclidean space using the distance matrix in which all distances are conserved, and then principal component analysis is performed. MDS and Goodness-of-fit (GOF) [53] were carried out using the *cmdscale* package in R [52]. K-means clustering was performed using the *kmeans* function in base R. Due to the lack of previous characterizations of RSV T cell immunity clusters, the number of T cell immunity groups was determined using the optimized within-cluster sum of square (wss) with Elbows method [54].

Calculation of genetic hamming distance

Genetic hamming distance, which is defined as the number of bases by which two nucleotide sequences differ, was calculated by comparing the number of different bases between each sequence in the subsampled datasets. The reconstructed most recent common ancestor (TMRCA) sequences for each dataset (subsampled F and G protein sequences of subtype A and subtype B, respectively) were estimated using the program "Treetime" and were used as root in our analysis [45].

Acknowledgments:

Funding: This work was supported by the National Institute of Allergy and Infectious Diseases, a component of the NIH, Department of Health and Human Services, under contract 75N93019C00052 (J.C. and J.B.) and Contract No. 75N93021C00018 (NIAID Centers of Excellence for Influenza Research and Response, CEIRR) (J.B.). J.B and S.T. are funded in part from the Centers for Disease Control under Contract No. 75D30119C06826 and 75D30121C11990.

Competing interests: A.S.DeG. and L.M are both paid employees of EpiVax. Some of the epitope prediction tools used in this study were developed by EpiVax.

Data and materials availability: Accession number to RSV sequence in the paper are available in supplementary materials. Code to generate T epitope landscapes are deposited in GitHub https://github.com/JianiC/RSV_Epitope.

References

[1] J. E. Crowe Jr and J. V Williams, "Paramyxoviruses: Respiratory Syncytial Virus and Human Metapneumovirus," *Viral Infect. Humans Epidemiol. Control*, pp. 601–627, Feb. 2014, doi: 10.1007/978-1-4899-7448-8_26.

[2] W.-J. Lee, Y. Kim, D.-W. Kim, H. S. Lee, H. Y. Lee, and K. Kim, "Complete Genome

Sequence of Human Respiratory Syncytial Virus Genotype A with a 72-Nucleotide Duplication in the Attachment Protein G Gene," *J. Virol.*, vol. 86, no. 24, pp. 13810 LP – 13811, Dec. 2012, doi: 10.1128/JVI.02571-12.

[3] J. S. McLellan, W. C. Ray, and M. E. Peeples, "Structure and function of respiratory syncytial virus surface glycoproteins," *Curr. Top. Microbiol. Immunol.*, vol. 372, pp. 83–104, 2013, doi: 10.1007/978-3-642-38919-1_4.

[4] J. Lee, L. Klenow, E. M. Coyle, H. Golding, and S. Khurana, "Protective antigenic sites in respiratory syncytial virus G attachment protein outside the central conserved and cysteine noose domains," *PLoS Pathog.*, vol. 14, no. 8, pp. e1007262–e1007262, Aug. 2018, doi: 10.1371/journal.ppat.1007262.

[5] "Updated guidance for palivizumab prophylaxis among infants and young children at increased risk of hospitalization for respiratory syncytial virus infection.," *Pediatrics*, vol. 134, no. 2, pp. 415–420, Aug. 2014, doi: 10.1542/peds.2014-1665.

[6] "Palivizumab, a humanized respiratory syncytial virus monoclonal antibody, reduces hospitalization from respiratory syncytial virus infection in high-risk infants. The IMpact-RSV Study Group.," *Pediatrics*, vol. 102, no. 3 Pt 1, pp. 531–537, Sep. 1998.

[7] T. F. Schwarz *et al.*, "Three dose levels of a maternal respiratory syncytial virus vaccine candidate are well tolerated and immunogenic in a randomized trial in non-pregnant women.," *J. Infect. Dis.*, Jun. 2021, doi: 10.1093/infdis/jiab317.

[8] C. Biagi *et al.*, "Current State and Challenges in Developing Respiratory Syncytial Virus Vaccines," *Vaccines*, vol. 8, no. 4, p. 672, Nov. 2020, doi: 10.3390/vaccines8040672.

[9] H. W. Kim *et al.*, "Respiratory syncytial virus disease in infants despite prior administration of antigenic inactivated vaccine.," *Am. J. Epidemiol.*, vol. 89, no. 4, pp.

422–434, Apr. 1969, doi: 10.1093/oxfordjournals.aje.a120955.

[10] B. R. Murphy, A. V Sotnikov, L. A. Lawrence, S. M. Banks, and G. A. Prince, “Enhanced pulmonary histopathology is observed in cotton rats immunized with formalin-inactivated respiratory syncytial virus (RSV) or purified F glycoprotein and challenged with RSV 3-6 months after immunization.,” *Vaccine*, vol. 8, no. 5, pp. 497–502, Oct. 1990, doi: 10.1016/0264-410x(90)90253-i.

[11] A. M. Killikelly, M. Kanekiyo, and B. S. Graham, “Pre-fusion F is absent on the surface of formalin-inactivated respiratory syncytial virus,” *Sci. Rep.*, vol. 6, p. 34108, Sep. 2016, doi: 10.1038/srep34108.

[12] F. P. Polack *et al.*, “A role for immune complexes in enhanced respiratory syncytial virus disease.,” *J. Exp. Med.*, vol. 196, no. 6, pp. 859–865, Sep. 2002, doi: 10.1084/jem.20020781.

[13] M. Connors, N. A. Giese, A. B. Kulkarni, C. Y. Firestone, H. C. 3rd Morse, and B. R. Murphy, “Enhanced pulmonary histopathology induced by respiratory syncytial virus (RSV) challenge of formalin-inactivated RSV-immunized BALB/c mice is abrogated by depletion of interleukin-4 (IL-4) and IL-10.,” *J. Virol.*, vol. 68, no. 8, pp. 5321–5325, Aug. 1994, doi: 10.1128/JVI.68.8.5321-5325.1994.

[14] M. E. Waris, C. Tsou, D. D. Erdman, S. R. Zaki, and L. J. Anderson, “Respiratory synctial virus infection in BALB/c mice previously immunized with formalin-inactivated virus induces enhanced pulmonary inflammatory response with a predominant Th2-like cytokine pattern.,” *J. Virol.*, vol. 70, no. 5, pp. 2852–2860, May 1996, doi: 10.1128/JVI.70.5.2852-2860.1996.

[15] N. I. Mazur *et al.*, “The respiratory syncytial virus vaccine landscape: lessons from the

graveyard and promising candidates.," *Lancet. Infect. Dis.*, vol. 18, no. 10, pp. e295–e311, Oct. 2018, doi: 10.1016/S1473-3099(18)30292-5.

[16] B. N. Blunck, W. Rezende, and P. A. Piedra, "Profile of respiratory syncytial virus prefusogenic fusion protein nanoparticle vaccine.," *Expert Rev. Vaccines*, vol. 20, no. 4, pp. 351–364, Apr. 2021, doi: 10.1080/14760584.2021.1903877.

[17] X. Liang *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).*, vol. 98, no. 19, 2019.

[18] A. Ahmed *et al.*, "Co-Circulation of 72bp Duplication Group A and 60bp Duplication Group B Respiratory Syncytial Virus (RSV) Strains in Riyadh, Saudi Arabia during 2014," *PLoS One*, vol. 11, no. 11, pp. e0166145–e0166145, Nov. 2016, doi: 10.1371/journal.pone.0166145.

[19] D. Tian *et al.*, "Structural basis of respiratory syncytial virus subtype-dependent neutralization by an antibody targeting the fusion glycoprotein," *Nat. Commun.*, vol. 8, no. 1, p. 1877, 2017, doi: 10.1038/s41467-017-01858-w.

[20] K. Hashimoto and M. Hosoya, "Neutralizing epitopes of RSV and palivizumab resistance in Japan," *Fukushima J. Med. Sci.*, vol. 63, no. 3, pp. 127–134, Dec. 2017, doi: 10.5387/fms.2017-09.

[21] W. M. Sullender, "Respiratory syncytial virus genetic and antigenic diversity," *Clin. Microbiol. Rev.*, vol. 13, no. 1, pp. 1–15, Jan. 2000, doi: 10.1128/CMR.13.1.1.

[22] X. Chen *et al.*, "Genetic variations in the fusion protein of respiratory syncytial virus isolated from children hospitalized with community-acquired pneumonia in China," *Sci. Rep.*, vol. 8, no. 1, p. 4491, 2018, doi: 10.1038/s41598-018-22826-4.

[23] G. Petrova, A. Ferrante, and J. Gorski, “Cross-reactivity of T cells and its role in the immune system,” *Crit. Rev. Immunol.*, vol. 32, no. 4, pp. 349–372, 2012, doi: 10.1615/critrevimmunol.v32.i4.50.

[24] A. S. De Groot *et al.*, “Better Epitope Discovery, Precision Immune Engineering, and Accelerated Vaccine Design Using Immunoinformatics Tools ,” *Frontiers in Immunology* , vol. 11. p. 442, 2020.

[25] L. He, A. S. De Groot, A. H. Gutierrez, W. D. Martin, L. Moise, and C. Bailey-Kellogg, “Integrated assessment of predicted MHC binding and cross-conservation with self reveals patterns of viral camouflage.,” *BMC Bioinformatics*, vol. 15 Suppl 4, no. Suppl 4, p. S1, 2014, doi: 10.1186/1471-2105-15-S4-S1.

[26] B. T. Grenfell, “Unifying the Epidemiological and Evolutionary Dynamics of Pathogens,” *Science (80-.).*, vol. 303, no. 5656, pp. 327–332, Jan. 2004, doi: 10.1126/science.1090727.

[27] S. D. J. *et al.*, “Mapping the Antigenic and Genetic Evolution of Influenza Virus,” *Science (80-.).*, vol. 305, no. 5682, pp. 371–376, Jul. 2004, doi: 10.1126/science.1097211.

[28] T. Bedford *et al.*, “Integrating influenza antigenic dynamics with molecular evolution,” *Elife*, vol. 3, Feb. 2014, doi: 10.7554/eLife.01914.

[29] B. Korber, M. LaBute, and K. Yusim, “Immunoinformatics comes of age,” *PLoS Comput. Biol.*, vol. 2, no. 6, pp. e71–e71, Jun. 2006, doi: 10.1371/journal.pcbi.0020071.

[30] B. Reynisson, B. Alvarez, S. Paul, B. Peters, and M. Nielsen, “NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data,” *Nucleic Acids Res.*, vol. 48, no. W1, pp. W449–W454, Jul. 2020, doi: 10.1093/nar/gkaa379.

[31] C. D. Russell, S. A. Unger, M. Walton, and J. Schwarze, “The Human Immune Response to Respiratory Syncytial Virus Infection.,” *Clin. Microbiol. Rev.*, vol. 30, no. 2, pp. 481–502, Apr. 2017, doi: 10.1128/CMR.00090-16.

[32] R. G. Woolthuis, C. H. van Dorp, C. Keşmir, R. J. de Boer, and M. van Boven, “Long-term adaptation of the influenza A virus by escaping cytotoxic T-cell recognition,” *Sci. Rep.*, vol. 6, no. 1, p. 33334, 2016, doi: 10.1038/srep33334.

[33] L. C. Katzelnick *et al.*, “Dengue viruses cluster antigenically but not as discrete serotypes,” *Science*, vol. 349, no. 6254, pp. 1338–1343, Sep. 2015, doi: 10.1126/science.aac5017.

[34] “Correction: Antigenic cartography of immune responses to Plasmodium falciparum erythrocyte membrane protein 1 (PfEMP1),” *PLOS Pathog.*, vol. 15, no. 8, p. e1008018, Aug. 2019.

[35] D. J. Smith *et al.*, “Mapping the antigenic and genetic evolution of influenza virus,” *Science (80-.).*, vol. 305, no. 5682, pp. 371–376, Jul. 2004, doi: 10.1126/science.1097211.

[36] J. Liu, T. J. Ruckwardt, M. Chen, T. R. Johnson, and B. S. Graham, “Characterization of respiratory syncytial virus M- and M2-specific CD4 T cells in a murine model,” *J. Virol.*, vol. 83, no. 10, pp. 4934–4941, May 2009, doi: 10.1128/JVI.02140-08.

[37] C. Viboud *et al.*, “Beyond clinical trials: Evolutionary and epidemiological considerations for development of a universal influenza vaccine,” *PLOS Pathog.*, vol. 16, no. 9, p. e1008583, Sep. 2020.

[38] J. Chen *et al.*, “Novel and extendable genotyping system for Human Respiratory Syncytial Virus based on whole-genome sequence analysis,” *Authorea Prepr.*, Jun. 2021, doi: 10.22541/AU.162251650.01202619/V1.

[39] S. S. Shepard, C. T. Davis, J. Bahl, P. Rivailler, I. A. York, and R. O. Donis, “LABEL: Fast and Accurate Lineage Assignment with Assessment of H5N1 and H9N2 Influenza A Hemagglutinins,” *PLoS One*, vol. 9, no. 1, p. e86921, Jan. 2014.

[40] “WHO | World Health Statistics 2011,” *WHO*, 2011.

[41] J. Rozewicki, S. Li, K. M. Amada, D. M. Standley, and K. Katoh, “MAFFT-DASH: integrated protein sequence and structural alignment,” *Nucleic Acids Res.*, vol. 47, no. W1, pp. W5–W10, Jul. 2019, doi: 10.1093/nar/gkz342.

[42] P. Rice, I. Longden, and A. Bleasby, “EMBOSS: The European Molecular Biology Open Software Suite,” *Trends Genet.*, vol. 16, no. 6, pp. 276–277, Jun. 2000, doi: 10.1016/S0168-9525(00)02024-2.

[43] S. S. Whitehead *et al.*, “Replacement of the F and G Proteins of Respiratory Syncytial Virus (RSV) Subgroup A with Those of Subgroup B Generates Chimeric Live Attenuated RSV Subgroup B Vaccine Candidates,” *J. Virol.*, vol. 73, no. 12, pp. 9773 LP – 9780, Dec. 1999, doi: 10.1128/JVI.73.12.9773-9780.1999.

[44] A. Stamatakis, “RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies,” *Bioinforma. Appl.*, vol. 30, no. 9, pp. 1312–1313, 2014, doi: 10.1093/bioinformatics/btu033.

[45] P. Sagulenko, V. Puller, and R. A. Neher, “TreeTime: Maximum-likelihood phylodynamic analysis,” *Virus Evol.*, vol. 4, no. 1, Jan. 2018, doi: 10.1093/VE/VEX042.

[46] G. Yu, D. K. Smith, H. Zhu, Y. Guan, and T. T. Y. Lam, “ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data,” *Methods Ecol. Evol.*, vol. 8, no. 1, pp. 28–36, Jan. 2017, doi: 10.1111/2041-210X.12628.

[47] L. Moise *et al.*, “iVAX: An integrated toolkit for the selection and optimization of antigens and the design of epitope-driven vaccines,” *Hum. Vaccin. Immunother.*, vol. 11, no. 9, pp. 2312–2321, 2015, doi: 10.1080/21645515.2015.1061159.

[48] Q. Zhu *et al.*, “A highly potent extended half-life antibody as a potential RSV vaccine surrogate for all infants,” *Sci. Transl. Med.*, vol. 9, no. 388, p. eaaj1928, May 2017, doi: 10.1126/scitranslmed.aaj1928.

[49] J. S. McLellan, Y. Yang, B. S. Graham, and P. D. Kwong, “Structure of respiratory syncytial virus fusion glycoprotein in the postfusion conformation reveals preservation of neutralizing epitopes,” *J. Virol.*, vol. 85, no. 15, pp. 7788–7796, Aug. 2011, doi: 10.1128/JVI.00555-11.

[50] A. H. Gutiérrez *et al.*, “T-cell epitope content comparison (EpiCC) of swine H1 influenza A virus hemagglutinin,” *Influenza Other Respi. Viruses*, vol. 11, no. 6, pp. 531–542, Nov. 2017, doi: 10.1111/irv.12513.

[51] E. Karosiene, C. Lundegaard, O. Lund, and M. Nielsen, “NetMHCcons: a consensus method for the major histocompatibility complex class I predictions,” *Immunogenetics*, vol. 64, no. 3, pp. 177–186, 2012, doi: 10.1007/s00251-011-0579-8.

[52] J. C. GOWER, “Some distance properties of latent root and vector methods used in multivariate analysis,” *Biometrika*, vol. 53, no. 3–4, pp. 325–338, Dec. 1966, doi: 10.1093/biomet/53.3-4.325.

[53] J. Sánchez, “MARDIA, K. V., J. T. KENT, J. M. BIBBY: Multivariate Analysis. Academic Press, London-New York-Toronto-Sydney-San Francisco 1979. xv, 518 pp., $ 61.00,” *Biometrical J.*, vol. 24, no. 5, p. 502, Jan. 1982, doi: https://doi.org/10.1002/bimj.4710240520.

[54] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set," *J. Stat. Software; Vol 1, Issue 6* , 2014, doi: 10.18637/jss.v061.i06.

Supporting information

S1 Fig. Distribution and diversity of T cell epitopes in RSV F protein. The tree panel on the left is a time-scaled phylogeny build with RSV-A **(A)** or RSV-B **(B)** F gene nucleotide sequences using the ML approach. Determined genotypes are labeled on the right with black bars. Each color column on the right side represents the presence of an MHC class I or class II epitope. Only the epitopes that are present in more than 1% of sampled isolates are displayed. The column color indicates different numbers of epitope sequences at the same location.

S2 Fig. Distribution and diversity of T cell epitopes in RSV G protein. The tree panel on the left is a time-scaled phylogeny build with RSV-A **(A)** or RSV-B **(B)** G gene nucleotide sequences using the ML approach. The clades that contain novel 72-nt or 60-nt duplication at the second hypervariable region of G gene were highlighted in red. Determined genotypes are labeled on the right with black bars. Each color column on the right side represents the presence of an MHC class I or class II epitope. Only the epitopes that are present in more than 1% of sampled isolates were displayed. The column color indicates different numbers of epitope sequences at the same location.

S3 Fig. Distribution of JanusMatrix Human Homology score for putative RSV MHC class I and class II epitopes. The cross-reactive potential of identified putative T cell epitopes and human host was represented with a JanusMatrix Human Homology score. 6.45% identified putative class I epitopes and 1.12% class II epitopes are cross-conserved on the TCR face with human epitopes.

S4 Fig. Predicted T cell epitope landscapes of RSV surface proteins. RSV T cell epitope landscapes were built with sequenced-based MHC class I epitope binding prediction (left), MHC class II epitope binding prediction (middle) or combining class I and class II epitope biding prediction (right). Sequences are colored by the epitope cluster determined by epitope landscapes built with combining Class I and Class II epitope prediction

S5 Fig. Total within sum of squares (*wss*) using *k-means* algorithm. Totals within sum of squares in epitope topographies were calculated after clustering into k (from 1 to 10) groups with *k-means*. The optimal number of clusters is determined to be 3 in the analysis of RSV-A F and G proteins and is determined to be 2 in the analysis of RSV-B F and G proteins using the Elbow method.

S6 Fig. Validation of T cell epitope distance estimation using the IEDB analysis resource. Validation is performed with MHC class I epitope binding prediction of RSV-A F protein. **(A)** Heatmaps for pairwise MHC class I epitope distance estimated in iVAX toolkits or calculated with custom python scripts using MHC class I molecule binding prediction that is implemented in IEDB. **(B)** Eigenvalues for each sequence are calculated from pairwise distance matrices using "RSpectra" package in R. The Pearson correlation test significantly supports a non-zero correlation between T cell epitope distance estimated with EpiCC and T cell epitope distance estimated with IEDB. **(C)** T cell epitope topographies are built with pairwise epitope distances estimated from EpiCC or IEDB. Both methods resulted in a similar cluster pattern for the CD8 T cell epitope profile of RSV-A F protein.

S7 Fig. Evaluation of RSV vaccine candidate strains with T cell epitope content in different WHO regions. RSV-A and RSV-B major surface protein sequences were grouped by isolation year and 6 isolated WHO regions, African Region (AFRO), Region of the Americas (PAHO), South-East Asia Region (SEARO), European Region (EURO), Eastern Mediterranean Region (EMRO) and Western Pacific Region (WPRO). Each year group was labeled by the latest isolated year of sequences after the previous group label. The proportion of cross-conserved T cell epitope content between vaccine strains (CP248 or CP52) and wild circulating strains in different year groups was represented by bar graphs.

S1 File. Accession number to RSV sequence that are used in this study.

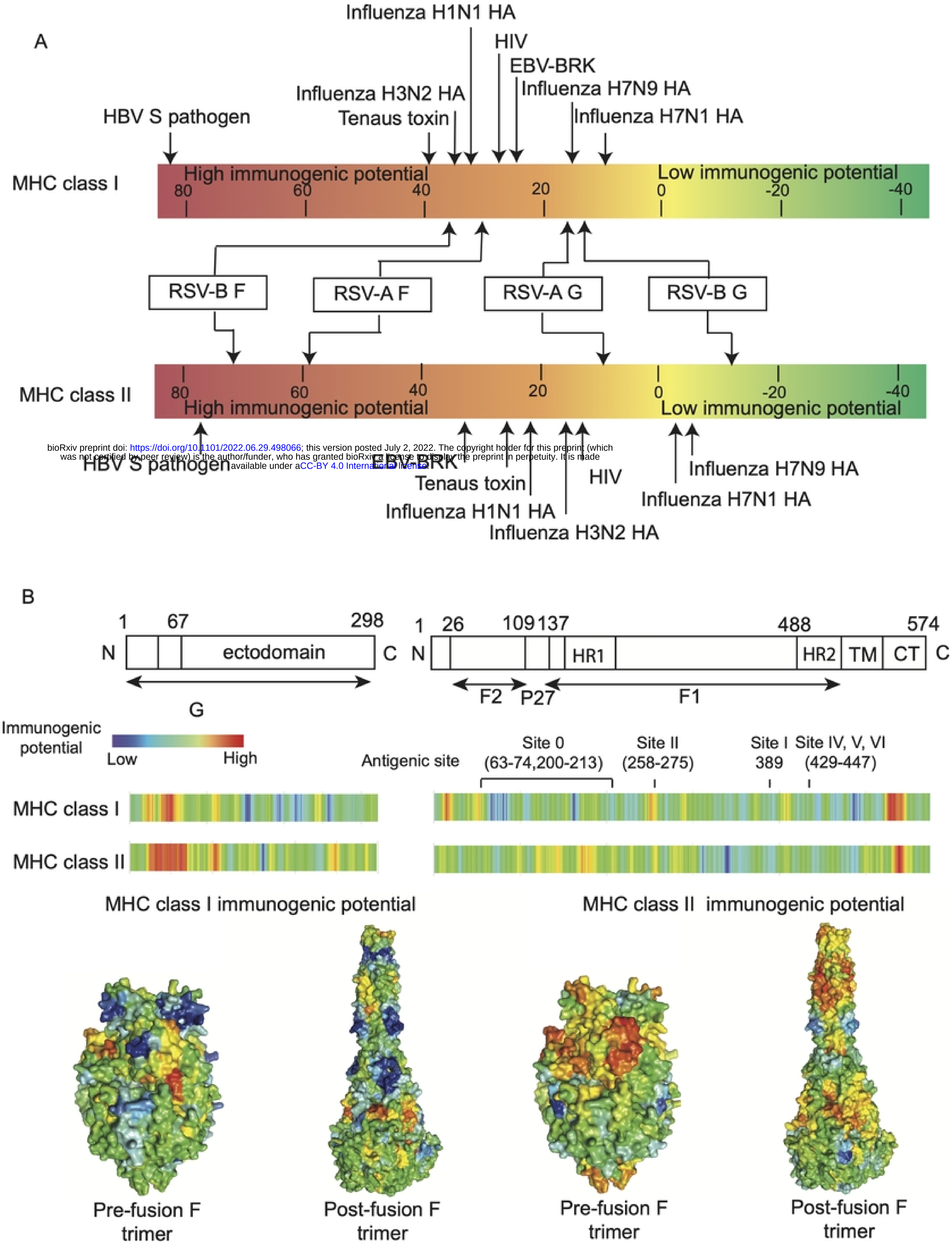

A
Influenza H1N1 HA
HIV
EBV-BRK
Influenza H7N9 HA
Influenza H3N2 HA
Influenza H7N1 HA
HBV S pathogen
Tenaus toxin
MHC class I
High immunogenic potential
Low immunogenic potential
80
60
40
20
0
-20
-40
RSV-B F
RSV-A F
RSV-A G
RSV-B G
MHC class II
80
60
40
20
0
-20
-40
High immunogenic potential
Low immunogenic potential
HBV S pathogen
EBV-BRK
Tenaus toxin
HIV
Influenza H7N9 HA
Influenza H1N1 HA
Influenza H7N1 HA
Influenza H3N2 HA
B
1
67
298
N
ectodomain
C
G
1
26
109
137
488
574
N
HR1
HR2
TM
CT
C
F2
P27
F1
Immunogenic potential
Low
High
Antigenic site
Site 0 (63-74,200-213)
Site II (258-275)
Site I 389
Site IV, V, VI (429-447)
MHC class I
MHC class II
MHC class I immunogenic potential
MHC class II immunogenic potential
Pre-fusion F trimer
Post-fusion F trimer
Pre-fusion F trimer
Post-fusion F trimer

Figure1

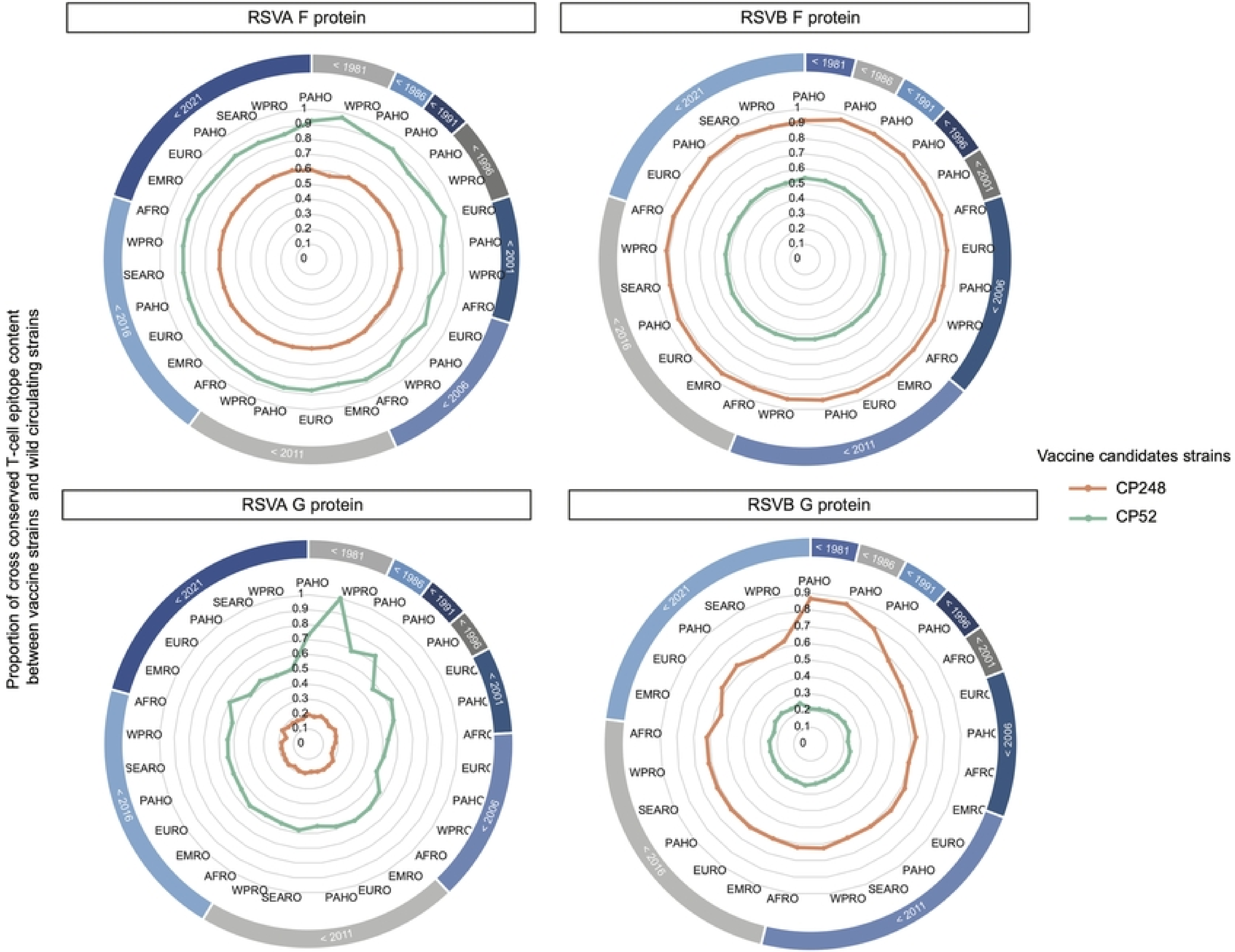

Proportion of cross conserved T-cell epitope content between vaccine strains and wild circulating strains
RSVA F protein
RSVB F protein
RSVA G protein
RSVB G protein
Vaccine candidates strains
CP248
CP52
< 1981
< 1986
< 1991
< 1996
< 2001
< 2006
< 2011
< 2016
< 2021
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
PAHO
WPRO
EURO
AFRO
EMRO
SEARO

Figure3

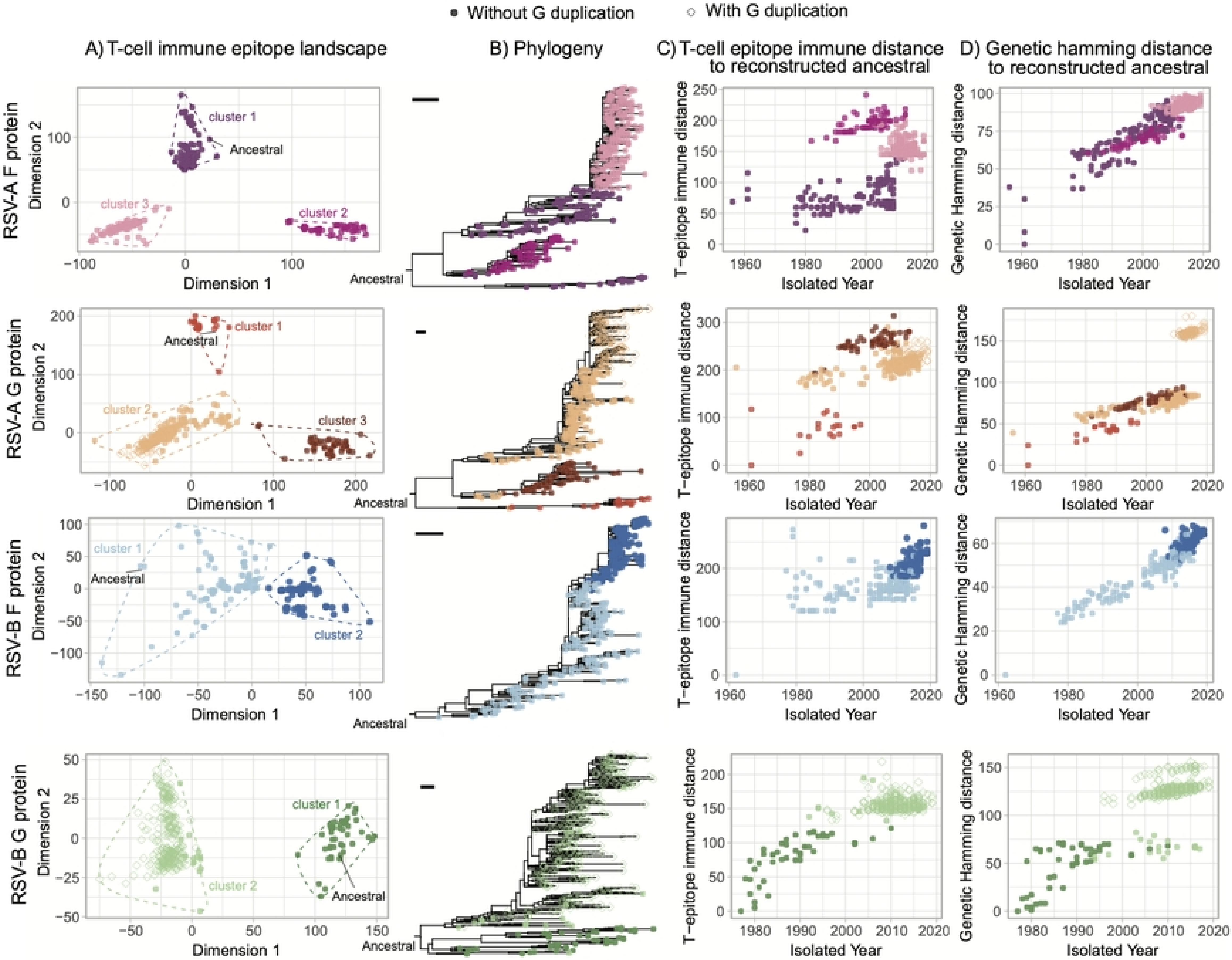

● Without G duplication
◇ With G duplication
A) T-cell immune epitope landscape
B) Phylogeny
C) T-cell epitope immune distance to reconstructed ancestral
D) Genetic hamming distance to reconstructed ancestral
RSV-A F protein
RSV-A G protein
RSV-B F protein
RSV-B G protein
cluster 1
cluster 2
cluster 3
Ancestral
Dimension 1
Dimension 2
T−epitope immune distance
Genetic Hamming distance
Isolated Year
-150
-100
-50
0
50
100
150
200
250
300
1960
1980
1990
2000
2010
2020
0
25
50
75
100
150
20
40
60

Figure2