Preprints.org

Article

Not peer-reviewed version

Phylogenomic Signatures of a Lineage of Vesicular Stomatitis Indiana Virus Circulating During the 2019–2020 Epidemic in the United States

Selene Zarate , Miranda R. Bertram , Case Rodgers , Kirsten Reed , Angela Pelzel-McCluskey , Ninnet Gomez-Romero , Luis L Rodriguez , Christie Mayo , Chad Mire , Sergei L. Kosakovsky Pond * , Lauro Velazquez-Salinas *

Posted Date: 8 October 2024

doi: 10.20944/preprints202410.0585.v1

Keywords: Vesicular stomatitis virus; evolution; positive selection; negative selection; epidemic lineages; natural selection

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Phylogenomic Signatures of a Lineage of Vesicular Stomatitis Indiana Virus Circulating during the 2019-2020 Epidemic in the United States

Selene Zarate [1], Miranda R. Bertram [2], Case Rodgers [3], Kirsten Reed [3], Angela Pelzel-McCluskey [4], Ninnet Gomez-Romero [5], Luis L. Rodriguez [2], Christie Mayo [3], Chad Mire [2], Sergei L. Kosakovsky Pond [6,*] and Lauro Velazquez-Salinas [2,*]

[1] Posgrado en Ciencias Genómicas, Universidad Autónoma de la Ciudad de Mexico, Mexico
[2] National Bio- and Agro-defense Facility, Agricultural Research Services, United States Department of Agriculture, Manhattan, NY, USA
[3] Department of Microbiology, Immunology and Pathology, Colorado State University, Fort Collins, CO, USA
[4] United States Department of Agriculture, Animal and Plant Health Inspection Service, Veterinary Services, Fort Collins, CO, USA
[5] Departamento de Microbiología e Inmunología, Facultad de Medicina Veterinaria y Zootecnia, Universidad Nacional Autónoma de México, Av. Universidad No. 3000 Col Copilco Universidad, Mexico City 14510, Mexico
[6] Institute for Genomics and Evolutionary Medicine, Department of Biology, Temple University, Philadelphia, PA 19122, USA
* Correspondence: spond@temple.edu (S.L.K.P.); lauro.velazquez@usda.gov (L.V.-S.)

Abstract: For the first time, we describe phylogenomic signatures of an epidemic lineage of vesicular stomatitis Indiana virus (VSIV). We applied multiple evolutionary analyses to a dataset of full-length genome sequences representing the circulation of an epidemic VSIV lineage in the US between 2019 and 2020. Based on phylogenetic analyses, we predicted the ancestral relationship of this lineage with a specific group of isolates circulating in the endemic zone of Chiapas, Mexico. Subsequently, our findings indicate that the lineage diversified into at least four different subpopulations during its circulation in the US. We identified single nucleotide polymorphisms (SNPs) that differentiate viral subpopulations and assessed their potential relevance using comparative phylogenetic methods, highlighting the preponderance of synonymous mutations during the differentiation of these populations. Purifying selection was the main evolutionary force favoring the conservation of this epidemic phenotype, with P and G genes as the main drivers of the evolution of this lineage. Our analyses identified multiple codon sites under positive selection and the association of these sites with specific functional domains at P, M, G, and L proteins. Based on ancestral reconstruction analyses, we showed the potential relevance of some of the sites identified under positive selection to the adaptation of the epidemic lineage at the population level. Finally, using a representative group of viruses from Colorado, we established a positive correlation between genetic and geographical distances, suggesting that positive selection on specific codon positions might have favored the adaptation of different subpopulations to circulation in specific geographical settings. Collectively, our study reveals the complex dynamics that accompany the evolution of an epidemic lineage of VSIV in nature. Our analytical framework provides a model for conducting future evolutionary analyses. The ultimate goal is to support the implementation of an early warning system for vesicular stomatitis virus in the US, enabling early detection of epidemic precursors from Mexico.

Keywords: vesicular stomatitis virus; evolution; positive selection; negative selection; epidemic lineages; natural selection.

© 2024 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

1. Introduction

Vesicular stomatitis virus (VSV) is an arbovirus [1] that belongs to the Rhabdoviridae family and the Vesiculovirus genus [2]. VSV is a negative-sense single-stranded RNA virus. VSV genome is approximately 11 kb long and contains five genes (N, P, M, G, L), which encode five structural proteins: nucleocapsid, phosphoprotein, matrix protein, glycoprotein, and large polymerase respectively [2]. Two main VSV serotypes have been described: vesicular stomatitis New Jersey virus (VSNJV) and vesicular stomatitis Indiana virus (VSIV), which cause numerous clinical cases in livestock in the Americas [3,4]. VSNJV is the more genetically diverse serotype, comprising six phylogenetic groups correlated with their geographical distribution [5]. In contrast, the VSIV serotype has fewer genetic clades, and the genetic differences between geographically dispersed isolates are lower than among the VSNJV isolates [6].

VSV can sporadically emerge from its endemic zones in southern Mexico and cause large epidemic outbreaks in the US [4,6–8]. The clinical manifestations of VSV and foot and mouth disease virus (FMDV) are indistinguishable; therefore, VSV outbreaks lead to the implementation of quarantines on infected premises until laboratory tests rule out the presence of FMDV [9]. Additionally, VSV contributes to significant economic losses associated with animal movement restrictions in the US [3,7,10,11].

Little is known about the factors associated with the emergence of epidemic VSV lineages in the US. Recent studies suggest that epidemic lineages of VSNJV represent a more virulent phenotype for vertebrate hosts than ancestral endemic lineages due to the increased ability to modulate the innate immune response [12]. Epidemic lineages of VSNJV also show an increased capacity to grow in insects in vivo [13]. Different outbreaks in the US have been consistently associated with the emergence of monophyletic lineages. However, since most of these outbreaks have been investigated using partial nucleotide sequences of the P gene [4,6–8]. As a result, the impact of variation in other genes on disease dynamics is unknown, and comprehensive analyses of natural selective forces cannot be carried out.

The two VSV serotypes differ in their capacity to emerge and sustain epidemic outbreaks in the US. During the last 30 years, VSNJV has emerged more frequently than VSIV [4,6–8]. However, in 2019 and 2020, an epidemic VSIV outbreak was registered in the US following a 21-year absence of this serotype [7]. This event represents the largest US VSV epidemic outbreak in the last 40 years. During 2019, a total of 1144 premises were affected in 111 counties and eight states (Colorado, Kansas, Nebraska, New Mexico, Oklahoma, Texas, Utah, and Wyoming), with the state of Colorado recording the most instances (693 premises in 38 counties). In 2020, a new outbreak affected 326 premises in 70 counties across eight states: Arizona, Arkansas, Kansas, Missouri, Nebraska, New Mexico, Oklahoma, and Texas [7].

To describe the evolutionary signatures of an epidemic lineage of VSIV in the US, we analyzed 98 full-length genome sequences from VSIV isolates collected from animals naturally infected between 2019 and 2020. To this end, we conducted a systematic evolutionary analysis using multiple high-resolution algorithms to detect natural selection. The findings from these analyses are discussed regarding the potential relevance of specific evolutionary patterns and critical genome sites that promoted this epidemic lineage's evolution during the US outbreak. Additionally, this knowledge may help implement surveillance programs to identify potential epidemic precursors in Mexico early.

2. Materials and Methods

2.1. Viral Sequences

The data set of this study included a total of 98 full-length genome sequences representing VSIV associated with the epidemic outbreak in the US during 2019-2020 (n=87) [14], previous isolates from the US (n=3), endemic zones of Mexico (n=2) [15], and Central and South America (n=6). These sequences were retrieved from GenBank. We aligned viral genomes using CLUSTAL W [16], as implemented in the BioEdit sequencing alignment editor.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

2.2. Phylogenetic Analysis

We inferred the maximum likelihood phylogeny of viral isolates in MEGA version 10.2.5 [17] using the GTR+G model (selected based on BIC) and assessed branching pattern support using 1000 bootstrap replicates. Analysis was conducted in MEGA version 10.2.5 [17].

2.3. Ancestral Sequence Reconstruction Analysis

The Maximum Likelihood method and General Time reversible mode inferred ancestral states at specific nucleotide sites. For each node, only the most likely nucleotide sequence is shown. Initial tree(s) for the heuristic search were obtained by Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood approach and selecting the topology with superior log likelihood value. Analyses were conducted in MEGA version 10.2.5 [17].

2.4. Pairwise Distance Analysis

We computed pairwise p-distances between VSIV isolates using MEGA version 10.2.5 [17] and obtained sampling variance estimates by bootstrap (1000 replicates).

2.5. Identification of Differential Single Polymorphic Sites (SNPs)

We determined SNPs that differentiate viral populations with the Metadata-driven Comparative analysis tool [18]. This algorithm performs a comparative analysis to identify positions in the genome that are significantly different among groups of sequences. A p-value of $5x10^{-6}$ was considered significant when accounting for multiple comparisons.

2.6. Population Structure Analysis

We used the fixation index test (FST) to evaluate the extent of genetic differentiation (population structure) between different phylogenetic groups of the epidemic VSIV lineage [19]. In this context, FST values may range between 0 and 1, indicating the existence of undifferentiated (panmictic) or structured populations, respectively. This analysis was conducted using HyPhy [20], and statistical significance for $FST \neq 0$ was assessed by a randomization test (1,000 replicates), and a p-value < 0.008 was considered significant to account for multiple comparisons.

2.7. Analysis of Molecular Variance (AMOVA)

Genetic distances were inferred using the package ape [21] ; subsequently, these distances were utilized to perform AMOVA calculation using the package pegas in R [22]. AMOVA was used to estimate potential population differentiation during the evolution of the epidemic lineage. AMOVA calculates the variance between and within groups, determining the level of divergence between each other in this way. One thousand permutations were carried out to assess the statistical significance of the differences.

2.8. Evolutionary Signatures

2.8.1. Identification Codons Evolving under Natural Selection

We sought to identify specific sites in the genome evolving under natural selection using a systematic approach as previously described for SARS-CoV-2 [23]. Multiple selection detection methods, including Fixed Effects Likelihood (FEL) [24] and Mixed Effects Model of Evolution (MEME) [25], were used. These methods detect sites under diversifying and purifying selection, acting in a pervasive (FEL) or episodic (MEME) manner by inferring rates of synonymous (dS) and nonsynonymous (dN) substitutions on a per-site basis in a codon-based phylogenetic framework [26], and conduct likelihood ratio statistical tests to assess deviations from neutrality ($dS = dN$). FEL was used to identify codons under pervasive diversifying ($dN > dS$) and purifying selection ($dN < dS$), while MEME was used to detect both pervasive and episodic diversifying selection [27].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

2.8.2. Assessing the Strength of Natural Selection during the Evolution of the Epidemic Lineage

We used the RELAX test to evaluate the relative strength of natural selection during the evolution of the epidemic VSIV lineage [28]. RELAX is a general hypothesis testing approach based on a codon-based phylogenetic framework to compare the distributions of dN/dS or ω (and thus the selective regimes) between two non-overlapping sets of branches in a tree. The intensification/relaxation parameter K, which maps ω→ω^K, determines whether there is evidence of relaxation (0<K<1, everything is shrunk towards ω = 1, or neutrality) or intensification (K>1, everything is pushed further away from 1) in the test set of branches relative to the reference set.

2.8.3. Recombination

The potential role of recombination during the evolution of the epidemic VSIV lineage was evaluated using GARD (Genetic Algorithm for Recombination Detection) [29]. This algorithm searches for the number and location of putative recombination breakpoints, which can cause potential topological incongruences in the phylogeny for different alignment parts. Differences in topology among different segments are evaluated by the posterior incongruence test (SH test) [30]. Evolutionary analyses were carried out using HyPhy v 2.5.52 [31] or later or the Datamonkey 2.0 web server [26].

2.9. Geographical Analysis

The correlation between geographical and pairwise genetic distances among different sequences obtained from Colorado was determined by the coefficient of determination (R2) analysis. For this purpose, pairwise genetic distances were calculated as described above, while hierarchical cluster analysis (HCA) was used to obtain a matrix of geographical distances among isolates. A p-value <0.05 was considered significant for R2 analysis. Additionally, the matrix of distances was assessed by analysis of variance (ANOVA) along with Tukey's honest significance test to estimate the number of geographical zones in Colorado. Analyses were conducted using JMP ® Pro version 16.0.0.

3. Results

3.1. The Genetic Origin of the Epidemic Lineage 2019-2020 Is Strongly Associated with VSIV Strains Circulating in Mexico's Endemic Zones

Phylogenetic analysis indicates that the 2019-2020 VSIV epidemic lineage is a conserved monophyletic lineage associated with the North American Clade (NA), with an average nucleotide identity within the lineage of 99.88% (Figure 1A and 1B). Based on the inferred topology, this lineage shares the most recent common ancestor and has a high degree of nucleotide similarity with the endemic Mexican VSIV isolates IN0817CPB and IN1017CPB. These isolates were recovered from cattle in the endemic region of Chiapas in 2017. Based on the topology and the lower levels of nucleotide identity predicted between the epidemic VSIV lineage 2019-2020 and the viral strains CO97E/97 and 98COE associated with the VSIV outbreak in the US in 1997-1998, we can suggest that the two epidemic events are independent (Figure 1A and 1B).

Interestingly, a 14 (AATTTTTTAATTTT) and 22 (AATTTTTTAATTTTAATTTT) nucleotide insertion in the intergenic noncoding region between genes G and L was also diagnostic of the epidemic lineage. The 22-nucleotide insertion was found in a minority of the isolates, including IN0819COE3, IN0819COE7, IN0819COE8, IN0819COE15, IN0819COE26, IN0919COE5 and IN1019COE2. The 14 insertion was also found in two strains from Mexico (IN0817CPB and IN1017CPB), supporting the ancestral relationship between them and the epidemic lineage. Conversely, this insertion is absent in previously epidemic VISV strains from the US (CO97E/97 and 98COE), implying two independent introductions in the US. Unexpectedly, we detected insertions in different intergenic regions associated with VSV's highly conserved seven uracil polyadenylation signals [32]. In this context, a polyadenylation signal of eight uracil (U8) was found in the intergenic N-P region of the ancestral strain IN0817CPB. A similar insertion was also present in the central American strain ES85B/L. Two different genotypes were found in the M-G intergenic region. On the

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

one hand, a U8 phenotype was found in the strain IN719COE11. On the other hand, a genotype carrying nine uracil in the polyadenylation signal (U9) was found in all epidemic strains recovered during 2020, including some strains recovered during 2019 (Wyoming/219838, Colorado/19559/2019 and Texas/ 18188/2019). Finally, a U8 genotype was found in the G-L intergenic region of the ancestral and epidemic strains IN1017CPB and IN719COE9, respectively.

Interestingly, a comparison between the consensus nucleotide sequence of the epidemic lineage 2019-2020 and the consensus sequence obtained from the ancestral endemic strains IN0817CPB and IN1017CPB identified a highly conserved mutation (CGG(R)-CAG(Q)) at gene L codon 1784 in all viruses recovered from the epidemic lineage.

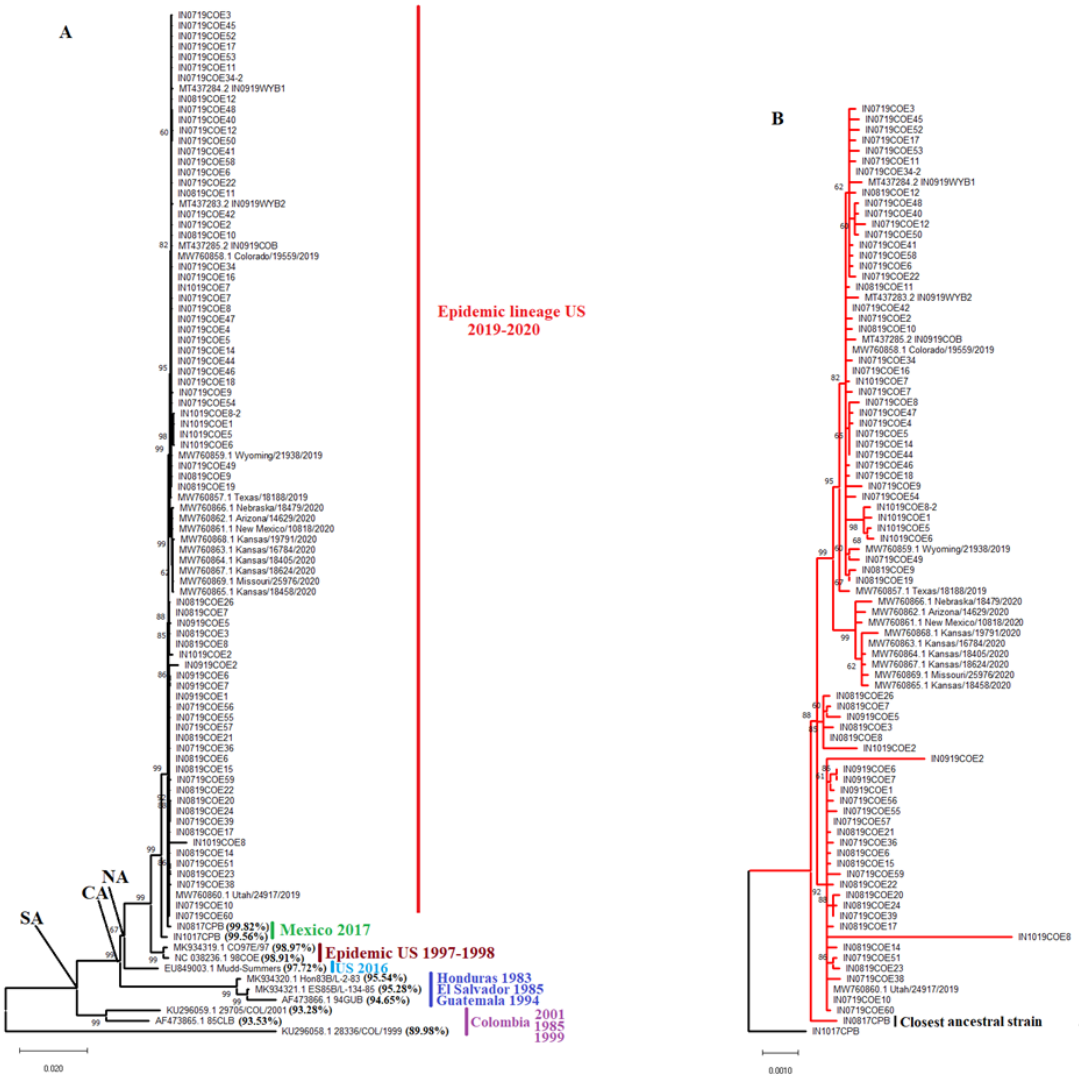


Figure 1. Identifying the ancestral relationship of the VSIV Epidemic lineage in the US. A) a maximum likelihood tree inferred using full-length genomic VSIV sequences, and the relationship between the epidemic VSIV lineage circulating in the USA during 2019-2020 and multiple earlier isolates from GenBank is shown. Branches are labeled with bootstrap support values. NA: North America, CA: Central America, SA: South America. Percentages in parentheses represent the average pairwise nucleotide identity between epidemic lineage sequences and the corresponding older isolate. B) Closeup from the phylogenetic analysis showing the ancestral relationship between the epidemic lineage and isolates from Chiapas, Mexico, IN0817CPB and IN1017CPB.

3.2. Epidemic VSIV 2019-2020 Lineage Diversified into Four Distinct Subpopulations in the US

Within the epidemic VSIV lineage, 2019-2020 are two main phylogenetic clades, from which at least four distinct genetic groups can be identified (Figure 2A). The segregation into four groups is

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

evidenced by high bootstrap values (84-100) for the corresponding internal tree branches and the results of an FST analysis (Figure 2B). Therefore, this lineage diversified into multiple subpopulations during its spread in the US. Additionally, population diversification of this epidemic lineage was confirmed for the analysis of molecular variance (AMOVA) (p-value 0.015).

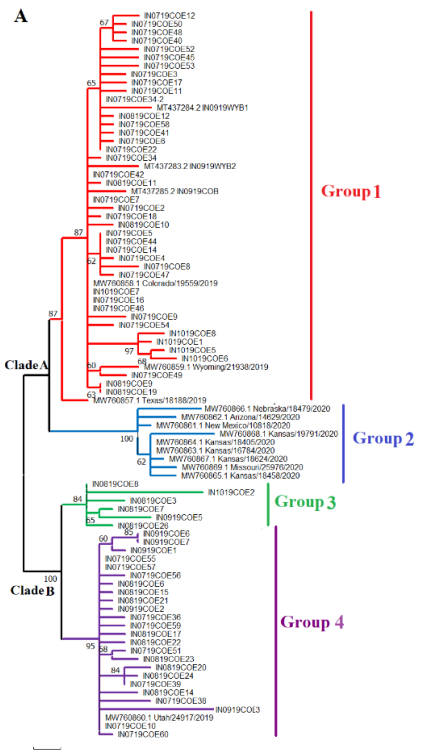
Clade A gave rise to two groups, I and II. Group I comprises a total of 42 viral sequences from Colorado obtained from horses between July and September of 2019, three sequences from Wyoming, one recovered from a horse (Wyoming/21938/2019), two from bovines (IN0919WYB1, IN0919WYB2) between July and September of 2019, and one viral sequence from Texas (Texas/18188/2019), recovered from a horse in June of 2019 (Figure 3A). The basal position of Texas/18188/2019 concerning the main cluster A indicates the ancestral relationship between this virus and viruses circulating in Colorado and Wyoming. Group II included nine horse isolates recovered from Arizona, Kansas, Missouri, Nebraska, and New Mexico, and it was the dominant lineage in 2020. Clade B comprises genetic groups III and IV. Group III includes a total of six horse isolates recovered from Colorado between August and September of 2019, while group IV contains viral sequences collected from horses in Colorado (n=24) and Utah (n=1) between July and September of 2019. Interestingly, all viral isolates showing the 22-nucleotide insertion in the G-L intergenic region were comprised in group III. No evidence of the circulation of viral lineages associated with groups I, III, and IV has been detected since 2019.

Furthermore, we identified 25 SNPs associated with the epidemic lineage's diversification into four viral subpopulations (Figure 3). P, G, and L genes accounted for most of the SNPs found in the genome. A total of 19 out of 25 SNPs were associated with synonymous mutations, while just 6 SNPs were producing nonsynonymous changes. Interestingly, 50% of the nonsynonymous changes were located at the P gene at codons 112, 161, and 239 (Figure 3), showing the potential relevance of this gene in the diversification of this epidemic lineage. To get more insights into the significance of these SNPs in the diversification of different genetic groups, we conducted an ancestral sequence reconstruction analysis. The results were consistent with a pattern of focal evolution at specific internal nodes followed by a period of conservation preserved across the leaf branches, indicating that these SNPs survived multiple transmission events and may promote the adaptation of the epidemic lineage at the population level (Figure S1, Appendix A). Five SNPs at codons P-112, P-161, G-439, G-455, and L-689 were associated with the diversification of this lineage into two main clades. On the other hand, four SNPs were identified at internal nodes of G1 (P-239, G-381, L-1653, and L-2021), eight at G2 (M-48, G-211, G-226, G-352, L-1707, L-1913, L-1955 and L-2048), three at G3 (P-116, G-263 and L-574) and four at G4 (P-13, L-863, L-975 and L-1006). Interestingly, some synonymous mutations at codons P-13, G-381, L-574, L-689, L-863, L-1913-, L-1955, and L-2021 were tracked in previous isolates from Central and South America origin. In the case of nonsynonymous mutations, just L-2048 was found in a lineage from South America (Figure S1, Appendix A).

There was no statistical evidence of relaxation or intensification of the strength of selection among different groups (RELAX test), suggesting that different groups evolved under evolutionary constraints indistinguishable using the available data.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

A



B

Comparison groups	Metapopulation diversity	Mean subpopulation diversity	Mean interpopulation diversity	F_{ST} 95% CI	p-value
Group 1 vs Group 2	0.00072	0.00043	0.00152	0.675-0.769	<0.0001
Group 1 vs Group 3	0.00091	0.00042	0.00151	0.690 -0.758	<0.0001
Group 1 vs Group 4	0.00068	0.00043	0.00166	0.700 -0.798	<0.0001
Group 2 vs Group 3	0.00103	0.00039	0.00203	0.767-0.857	<0.0001
Group 2 vs Group 4	0.00133	0.0005	0.00223	0.750 -0.856	<0.0001
Group 3 vs Group 4	0.0006	0.0004	0.00112	0.520 -0.778	<0.0001

Figure 2. Population structure of the VSIV epidemic lineage 2019-2020 in the US. A) A phylogenetic analysis was conducted by maximum likelihood to show the main events of diversification in the epidemic lineage during its circulation in the US. B) Fixation index test (FST) analysis supporting the existence of four divergent groups..

Gene	Codon	Position	Change	Chi-square Value	P-value	G1	G2	G3	G4
P	13	39	Synonymous	87.00	6.511E-15	$_{(S)}$teC$_{(47)}$	$_{(S)}$teC$_{(9)}$	$_{(S)}$teC$_{(6)}$	$_{(S)}$teT$_{(25)}$
P	112	334	Nonsynonymous	87.00	6.51E-15	$_{(K)}$Aag$_{(47)}$	$_{(K)}$Aag$_{(9)}$	$_{(E)}$Gag$_{(6)}$	$_{(E)}$Gag$_{(25)}$
P	116	348	Synonymous	86.93	6.702E-15	$_{(D)}$gaC$_{(47)}$	$_{(D)}$gaC$_{(9)}$	$_{(D)}$gaT$_{(6)}$	$_{(D)}$gaC$_{(25)}$
P	161	481	Nonsynonymous	87.00	6.51E-15	$_{(T)}$Aca$_{(47)}$	$_{(T)}$Aca$_{(9)}$	$_{(A)}$Gca$_{(6)}$	$_{(A)}$Gca$_{(25)}$
P	239	716	Nonsynonymous	83.07	3.959E-14	$_{(L)}$cTa$_{(46)}$-$_{(R)}$cGa$_{(1)}$	$_{(R)}$cGa$_{(9)}$	$_{(R)}$cGa$_{(6)}$	$_{(R)}$cGa$_{(25)}$
M	48	144	Synonymous	86.96	6.614E-15	$_{(V)}$gtT$_{(47)}$	$_{(V)}$gtG$_{(9)}$	$_{(V)}$gtT$_{(6)}$	$_{(V)}$gtT$_{(25)}$
G	211	631	Synonymous	86.96	6.614E-15	$_{(L)}$Cta$_{(47)}$	$_{(L)}$Tta$_{(9)}$	$_{(L)}$Cta$_{(6)}$	$_{(L)}$Cta$_{(25)}$
G	226	678	Synonymous	86.96	6.614E-15	$_{(F)}$ttT$_{(47)}$	$_{(F)}$ttC$_{(9)}$	$_{(F)}$ttT$_{(6)}$	$_{(F)}$ttT$_{(25)}$
G	263	789	Synonymous	27.60	0.001114	$_{(A)}$gcA$_{(47)}$	$_{(A)}$gcA$_{(9)}$	$_{(A)}$gcA$_{(4)}$-$_{(A)}$gcG$_{(2)}$	$_{(A)}$gcA$_{(25)}$
G	352	1055	Nonsynonymous	86.96	6.614E-15	$_{(A)}$gCt$_{(47)}$	$_{(V)}$gTt$_{(9)}$	$_{(A)}$gCt$_{(6)}$	$_{(A)}$gCt$_{(25)}$
G	381	1143	Synonymous	87.00	6.504E-15	$_{(A)}$gaT$_{(47)}$	$_{(A)}$gaC$_{(9)}$	$_{(A)}$gaC$_{(6)}$	$_{(A)}$gaC$_{(25)}$
G	439	1317	Synonymous	87.00	6.51E-15	$_{(L)}$ttG$_{(47)}$	$_{(L)}$ttG$_{(9)}$	$_{(L)}$ttA$_{(6)}$	$_{(L)}$ttA$_{(25)}$
G	455	1365	Synonymous	87.00	6.51E-15	$_{(E)}$gaG$_{(47)}$	$_{(E)}$gaG$_{(9)}$	$_{(E)}$gaA$_{(6)}$	$_{(E)}$gaA$_{(25)}$
L	574	1722	Synonymous	86.93	6.702E-15	$_{(D)}$gaC$_{(47)}$	$_{(D)}$gaC$_{(9)}$	$_{(D)}$gaT$_{(6)}$	$_{(D)}$gaC$_{(25)}$
L	689	2067	Synonymous	87.00	6.51E-15	$_{(S)}$agC$_{(47)}$	$_{(S)}$agC$_{(9)}$	$_{(S)}$agT$_{(6)}$	$_{(S)}$agT$_{(25)}$
L	863	2589	Both	21.25	0.01157	$_{(D)}$gaT$_{(47)}$	$_{(D)}$gaT$_{(47)}$	$_{(D)}$gaT$_{(5)}$-$_{(E)}$gaA$_{(1)}$	$_{(D)}$gaT$_{(22)}$-$_{(D)}$gaC$_{(3)}$
L	975	2925	Synonymous	87.00	6.511E-15	$_{(E)}$gaA$_{(47)}$	$_{(E)}$gaA$_{(9)}$	$_{(E)}$gaA$_{(6)}$	$_{(E)}$gaG$_{(25)}$
L	1006	3018	Synonymous	87.00	6.511E-15	$_{(I)}$atA$_{(47)}$	$_{(I)}$atA$_{(9)}$	$_{(I)}$atA$_{(6)}$	$_{(I)}$atC$_{(25)}$
L	1653	4957	Synonymous	17.98	0.03539	$_{(R)}$Cgg$_{(30)}$-$_{(R)}$Agg$_{(17)}$	$_{(R)}$Cgg$_{(9)}$	$_{(R)}$Cgg$_{(6)}$	$_{(R)}$Cgg$_{(25)}$
L	1707	5121	Synonymous	86.96	6.614E-15	$_{(P)}$ccT$_{(47)}$	$_{(P)}$ccC$_{(9)}$	$_{(P)}$ccT$_{(6)}$	$_{(P)}$ccT$_{(25)}$
L	1913	5739	Synonymous	77.35	5.424E-13	$_{(S)}$tcT$_{(46)}$-$_{(S)}$tcC$_{(1)}$	$_{(S)}$tcC$_{(9)}$	$_{(S)}$tcT$_{(6)}$	$_{(S)}$tcT$_{(25)}$
L	1955	5865	Synonymous	86.96	6.614E-15	$_{(P)}$ccC$_{(47)}$	$_{(P)}$ccT$_{(9)}$	$_{(P)}$ccC$_{(6)}$	$_{(P)}$ccC$_{(25)}$
L	2021	6063	Synonymous	83.07	3.959E-14	$_{(P)}$ccG$_{(45)}$-$_{(P)}$ccA$_{(1)}$	$_{(P)}$ccA$_{(9)}$	$_{(P)}$ccA$_{(6)}$	$_{(P)}$ccA$_{(25)}$
L	2048	6143	Nonsynonymous	55.82	8.491E-09	$_{(R)}$cGt$_{(47)}$	$_{(R)}$cGt$_{(3)}$-$_{(H)}$cAt$_{(6)}$	$_{(R)}$cGt$_{(6)}$	$_{(R)}$cGt$_{(25)}$

Figure 3. Metadata-driven comparative analysis. The SNPs associated with specific codon sites that are significantly different from the null expectation among phylogenetic groups (p-value cutoff <0.05) were identified by the Metadata-driven comparative analysis. G1 to G4 columns represent the codon composition of different phylogenetic groups. Specific SNPs at each codon are highlighted in capital letters. The column position indicates the nucleotide position in the coding sequence at specific genes

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

where the SNP was identified. Parentheses on the left (right) indicate the amino acid encoded and the number of sequences associated with this codon at any specific group, respectively.

3.3. Episodic Diversifying Selection Is a Distinctive Evolutionary Hallmark of VSIV in Nature

We used codon phylogenetic models to quantify selection pressures on viral genes, using the maximum likelihood tree inferred from full-length genomes (Figure 1). Overall, all genes were subject to purifying selection, on average (Figure 4), with all estimates significantly different from dN/dS=1 (neutrality). Using a pairwise likelihood ratio test procedure at p≤0.05 with a Holm-Bonferroni correction [33,34], we obtained a partial ordering of genes based on their average level of conservation, shown in Figure 4. Genes N and M were the most conserved, while genes G and P – were the least conserved.

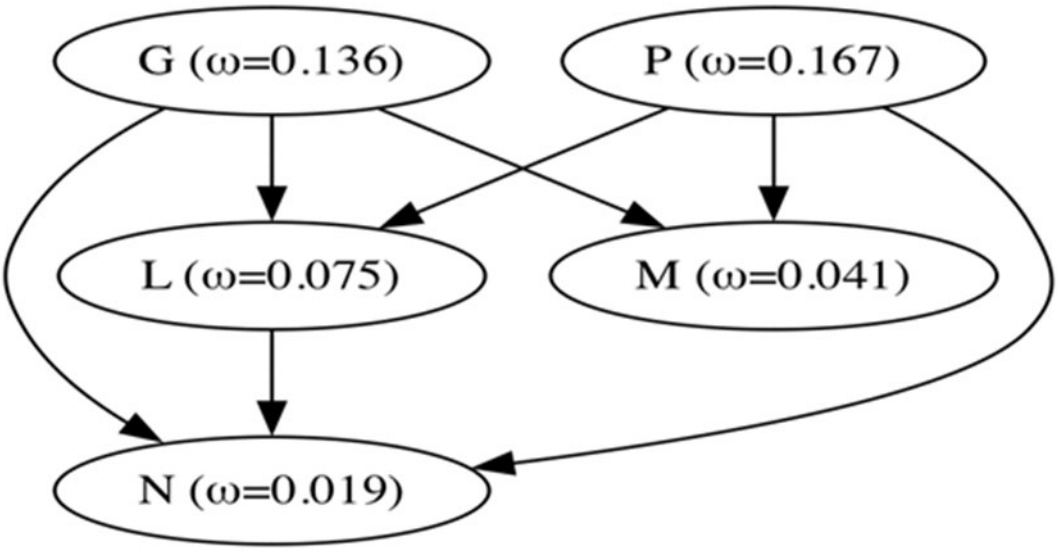


Figure 4. A partial ordering of VSV genes based on their average conservation (mean ω). A directed arrow between gene X and Y is a statement that ω (X) > ω (Y) with statistical significance (p ≤ 0.05).

Because of relatively low levels of diversity (tree lengths measured subs/site), we used versions of the FEL and MEME site-level detection methods, which are more suitable for the small sample, low diversity setting; they utilize parametric bootstrap with 100 replicates to assess significance [35]. Similar to the rankings based on comparing mean dN/dS estimates, genes N and M were the most conserved (largest fractions of negatively selected sites, smallest fractions of positively selected sites), while genes G and P were the least conserved. Overall, the more sensitive versions of FEL and MEME found evidence of purifying selection (at p ≤ 0.05) on 565 sites and of positive selection on 42 sites across the entire genome (Table 1). Specific sites under positive and purifying selection are shown in Figure 5 (Appendix B) and Figure S2 (Appendix A), respectively.

No breakpoints were detected by the GARD analysis, indicating that recombination did not play a significant role during the evolution of this sample of VISV genomes.

Table 1. Overview of evolutionary pressures across the genome of VSIV. A dN/dS at each gene was computed using the MG94xREV model. The confidence interval was estimated using profile likelihood.

Gene	Codons	Tree length	dS	dN	dN/dS (95% CI)	Sites under selection (p ≤ 0.05)	Sites under selection / 1000 codons

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

						Purifying	Positive	Purifying	Positive
N	422	0.176	0.814	0.011	0.019 (0.01-0.03)	60	1	142.2	2.4
P	265	0.259	0.882	0.092	0.167 (0.13-0.21)	37	4	139.6	15.1
M	229	0.238	1.062	0.032	0.041 (0.03-0.06)	45	2	196.5	8.7
G	511	0.263	0.933	0.089	0.136 (0.11-0.16)	81	11	158.5	21.5
L	2109	0.229	0.921	0.048	0.075 (0.07-0.09)	342	24	162.2	11.4
Genome	3536	0.230	0.912	0.051	0.081 (0.08-0.09)	565	42	159.8	11.9

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

GENE	Codon	α	β^1	p^1	β^+	p^+	LRT	p-value	# branches	Class
N	360	0.00	0.00	0.00	512.39	1.00	2.82	0.04	1	Diversifying
P	45	0.00	0.00	0.00	92.08	1.00	4.84	0.04	1	Diversifying
P	194	0.00	0.00	0.00	829.95	1.00	3.18	0.03	1	Diversifying
P	212	0.00	0.00	0.00	853.26	1.00	3.32	0.03	1	Diversifying
P	239	9.30	3.86	0.00	4244.01	1.00	2.81	0.03	1	Diversifying
M	30	0.00	0.00	0.00	303.93	1.00	2.51	0.02	1	Diversifying
M	33	0.00	0.00	0.00	1059.82	1.00	3.36	0.03	1	Diversifying
G	24	0.00	0.00	0.00	977.31	1.00	6.43	0.01	2	Diversifying
G	115	26.57	0.00	0.00	1288.38	1.00	2.17	0.05	1	Diversifying
G	217	0.00	0.00	0.00	71.21	1.00	3.64	0.04	2	Diversifying
G	234	0.00	0.00	0.00	790.40	1.00	3.19	0.05	1	Diversifying
G	241	0.00	0.00	0.00	627.75	1.00	2.78	0.05	1	Diversifying
G	248	19.59	0.00	0.00	2960.95	1.00	3.00	0.03	1	Diversifying
G	258	0.00	0.00	0.00	967.99	1.00	3.18	0.01	1	Diversifying
G	271	0.00	0.00	0.00	1101.65	1.00	6.11	0.01	2	Diversifying
G	274	0.00	0.00	0.00	649.79	1.00	2.77	0.05	1	Diversifying
G	403	14.65	0.00	0.00	1903.82	1.00	2.73	0.05	1	Diversifying
G	505	0.00	0.00	0.00	1816.88	1.00	3.70	0.04	1	Diversifying
L	58	0.00	0.00	0.00	632.93	1.00	2.73	0.05	1	Diversifying
L	68	0.01	0.00	0.00	219.05	1.00	3.10	0.04	1	Diversifying
L	122	0.00	0.00	0.00	1980.52	1.00	3.89	0.02	1	Diversifying
L	126	0.00	0.00	0.00	5269.75	1.00	5.28	0.01	1	Diversifying
L	174	33.58	0.00	0.00	6685.65	1.00	3.70	0.03	1	Diversifying
L	212	0.00	0.00	0.00	2916.29	1.00	4.34	0.02	1	Diversifying
L	266	0.00	0.00	0.00	7013.05	1.00	5.10	0.01	1	Diversifying
L	395	14.69	0.00	0.00	2539.27	1.00	3.15	0.03	1	Diversifying
L	644	0.00	0.00	0.00	1363.17	1.00	3.62	0.03	1	Diversifying
L	695	0.00	0.00	0.00	3138.55	1.00	4.52	0.01	1	Diversifying
L	1251	0.00	0.00	0.00	6487.61	1.00	4.87	0.01	1	Diversifying
L	1272	24.30	0.00	0.00	5196.29	1.00	3.78	0.04	1	Diversifying
L	1298	0.00	0.00	0.00	851.46	1.00	5.63	0.01	2	Diversifying
L	1345	0.00	0.00	0.00	7419.07	1.00	5.37	0.03	1	Diversifying
L	1590	84.15	0.00	0.00	7460.00	1.00	3.18	0.04	1	Diversifying
L	1591	29.04	0.00	0.00	7023.16	1.00	4.04	0.02	1	Diversifying
L	1605	48.39	0.00	0.00	7944.44	1.00	3.60	0.01	1	Diversifying
L	1622	0.00	0.00	0.00	807.89	1.00	3.66	0.03	1	Diversifying
L	1784	0.00	0.00	0.00	7653.44	1.00	6.22	0.01	1	Diversifying
L	1887	0.00	0.00	0.00	2846.81	1.00	4.35	0.02	1	Diversifying
L	1953	7.69	0.00	0.00	1078.69	1.00	2.64	0.04	1	Diversifying
L	2004	0.00	0.00	0.00	838.76	1.00	3.31	0.04	1	Diversifying
L	2006	0.00	0.00	0.00	1182.36	1.00	3.46	0.05	1	Diversifying
L	2045	0.01	0.00	0.00	9759.98	1.00	5.52	0.03	1	Diversifying

Figure 5. Identification of codon sites evolving under positive selection in natural populations of VSIV. The figure shows the 42 codons under positive selection identified at multiple genes VSIV by MEME analysis. a: synonymous substitution rate, b^1:Non-synonymous substitution rate for the negative /neutral evolution component 1, p^1: mixture distribution weight allocated to negative/neutral evolution component 1, b^+:non-synonymous substitution rate at a site for the positive selection component, p^+:mixture distribution weight allocated to the positive selection component, LTR: likelihood test statistics for episodic diversification, i.e., $p^+ > 0$, p-value: asymptotic p-value for episodic diversification, i.e., $p^+>0$, # branches: the (very approximate and rough)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

estimate of how many branches have been under selection at this site, i.e., had an empirical Bayes factor of 100 or more for the b^+ rate, q: and class: selection kind..

3.4. The Evolution of the Epidemic VSIV Lineage Is Constrained by the Functionality of Its Proteins

Once we determined specific codon sites at different genes under positive and purifying selection on VSIV populations in nature (Figure 5), we focused on understanding how these codon sites impacted the evolution of the epidemic lineage in the US. To comprehend if functional protein constraints might influence the location of codons under selection, we conducted an extensive literature review about previously reported functional sites at different VSIV proteins (Figure 6).

Based on dN/dS ratios and the percentage of invariable codon sites, N was the most conserved gene during the evolution of the epidemic lineage (Figure 6A). No codons were detected under positive selection in N. Overall, we observed high conservation in codons encoding critical residues associated with viral RNA interactions, N-M or N-P protein interaction, and N-self interactions (Figure 6A). A single nonsynonymous mutation affecting a unique isolate from group 4 (IN0819COE20) at codon 13 ($GTC_{(V)}$-$ATC_{(I)}$) was tracked on this gene, located at the N- terminal site of N protein specifically in the functional site N0-P, associated with a binding site that prevents nascent N molecules from self-assembling and from binding to cellular RNAs [36]. Three codons were found under purifying selection, none associated with a residue implied in a functional site.

P gene appeared as the second most divergent gene in the epidemic lineage (Figure 6B). However, nonsynonymous mutations accumulated at specific functional regions, and high conservation was seen in residues associated with P-N0 chaperone region, P-L interactions (polymerase cofactor, transcriptional activity), and P-N binding. Three nonsynonymous mutations (P-112, P-161, P-239) previously predicted by our Metadata-driven comparative analysis (potentially promoting the adaptation of this lineage at the population level) were tracked in this gene. Mutations at codons P-112 and P-161, located in domain I and hinge region (both parts of the autodimerization region of P protein) (Figure 6B), were categorized as neutral changes by MEME and FEL analyses. Conversely, codon P-239 (Figure 3) was under positive selection (Figure 5). The residue encoded by this codon is in domain II, close to codons P-233-235, encoding residues implicated in the binding between P and N proteins [37] (figures 5 and 6B).

Additionally, MEME and FEL detected codon sites P-194 (Hinge region) and 212 (domain II) under positive selection. The selection at these codon sites affected single isolates belonging to groups 2 and 1, respectively (Figure S3, Appendix A). A total of 6 codons were detected under purifying selection on this gene, preserving residues encoded by these codons at different functional sites on the P protein (Figure 6B).

The M gene showed high conservation during the evolution of the epidemic lineage in the US (Figure 6C). We noticed that sites associated with functional regions in the N-M complex, viral release, membrane association, M-G interactions, and M self-interactions displayed high overall conservation. However, two nonsynonymous mutations were predicted in this gene, affecting single isolates at two different functional sites. The first mutation at codon 14 ($GGT_{(G)}$-$AGT_{(S)}$) was found in the isolate IN0819COE12 (group 1) and was characterized as a neutral evolving change by MEME and FEL. Interestingly, this mutation is located in a previously predicted functional motif spanning residues 14-19 (Figure 6C), which impairs the release of viral particles from infected cells [38]. The second mutation was located at codon 33 on a single isolate of group 1 (Figure S3, Appendix A) and was predicted to be under positive selection (Figure 5). This mutation disrupts the highly conserved ATG codon, which is linked to the expression of M2, one of the three recognized forms of the M protein [39,40]. Additionally, three codons were predicted to be under purifying selection (Figure 6C).

The G gene was the most divergent in the epidemic lineage (Figure 6D), with mutations accumulated at specific locations along different functional domains (DI-DIV). High conservation was observed in the transmembrane domine (TM) and residues associated with G-M interactions. A total of six codons under positive selection were predicted on this gene (Figure 5). Three mutations at codons 24, 271, and 403 were located at DII (Trimerization domain). At codon 24, we detected

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

mutations in single isolates from genetic groups 1 and 3. However, in each isolate, the dominant codon CAC(H) was changed to either AAC (N) or CGC(R). Similarly, at codon 271, codon GAA(E) changed to GGA(G) or GAC(D) in single isolates from groups 1 and 4. Mutation at codon 403 was linked to a single viral isolate from group 2 (Figure S3, Appendix A). In domain III (DIII/pH domine), codons 234 and 248 were under positive selection. Mutations at these codons were found independently in single isolates from genetic groups 1 and 4. Finally, an additional site under positive selection was found in domain IV (DIV/fusion domain) at codon 115. The mutation at this codon was in a single isolate from group 1 (Figure S3, Appendix A). Interestingly, codons 115 and 271 were associated with residues recognized as neutralizing epitopes [41]. In this context, residue encoded by codon 352, located in Domain I (DI/lateral domain), was also recognized as a neutralizing epitope. This residue was previously identified through our Metadata-driven comparative analysis (linked with the emergence of G2) (Figure S1, Appendix A) and classified as a neutrally evolving site by our MEME and FEL analyses. A total of seven codons in this gene were identified under purifying selection (Figure 6D).

The gene L was the third most conserved gene during the evolution of the epidemic lineage (Figure 6E). A high level of conservation was observed in the conserved regions II, IV, and VI (CRII, CRIV, and CRVI). These regions contain critical motifs in multiple polymerase regions and critical residues involved in capping and methyltransferase activities. A total of 17 codons were predicted to be under positive selection at this gene (Figure 6E). Out of these, eight codons were located at the RNA polymerase domain (RdRp) at codon positions 122, 126, 174, 212, 266, 395, 644, and 695. These codons impact isolates from groups 1, 3, and 4 (Figure S3, Appendix A). Interestingly, mutations at some of these codons (266, 395, 644, and 695) disrupted residues in conserved regions I (CRI) and III (CRIII) (Figure 6E). In this sense, three codons (1251, 1272, and 1298) located at the capping domain (Cap) produced changes at residues of the conserved region V (CRV) (Figure 6E). These mutations were associated with individual isolates of groups 1 and 2 (Supplementary file 3). However, mutation at codon 1298 was also present in all VSIV isolates from Central and South American clades (Figure 1 and Figure S4, Appendix A). Notably, FEL analysis conducted on internal branches detected codon 1298 to be under positive selection (p-value= 0.07), indicating the potential of this codon site to induce adaptation at the population level on VSIV in nature.

Three codons (1345, 1590, and 1591) were predicted in linker regions under positive selection (Figure 6E). At codon 1345, mutations were found in two independent isolates from genetic group 1, while single isolates from groups 2 and 3 carried mutations at codons 1590 and 1591(Figure S3, Appendix A). Finally, the last four codons under positive selection were distributed between methyltransferase (MT) and C-terminal (CTD) domains (Figure 6D). The MT domain included codons 1605 and 1887, where three independent isolates from group 4 and one single isolate were identified as carrying mutations at these codons, respectively (Figure S3, Appendix A). Remarkably, in this domain, the codon L-1784 was predicted under positive selection (Figure 5). The L-1784 codon not only differentiates the endemic ancestral viruses (IN0817CPB and IN1017CPB) from the epidemic lineage by the mutation CGG(R)-CAG(Q) but also from the other VSIV previously reported in nature (Figure S4, Appendix A). This codon is part of the conserved region VI (CRVI) (Figure 6E), where CGG(R) was the dominant allele among VSIV in nature. However, the presence and conservation of allele CAG(Q) during the circulation of the epidemic lineage in the US stress the potential importance of this mutation in the emergence of this lineage.

The CTD domain included codons 2006 and 2045. Single isolates from groups 3 and 2 were identified with mutations at these codons (Figure S3, Appendix A). Moreover, CTD includes codon 2048 (Figure 6D), identified by our Metadata-driven comparative analysis (Figure 3) and implicated in the divergence within G2 (Figure S1, Appendix A). Despite this codon being classified as a neutrally evolving site by MEME and FEL analyses, mutation at this codon CGT(R)-CAT(H) present in all isolates recovered from Kansas and Missouri during the epidemic outbreak in 2020, was also found in the south American isolate 85CLB.

Regarding purifying selection, 30 codons distributed at different functional sites were predicted on the L gene (Figure 6E).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

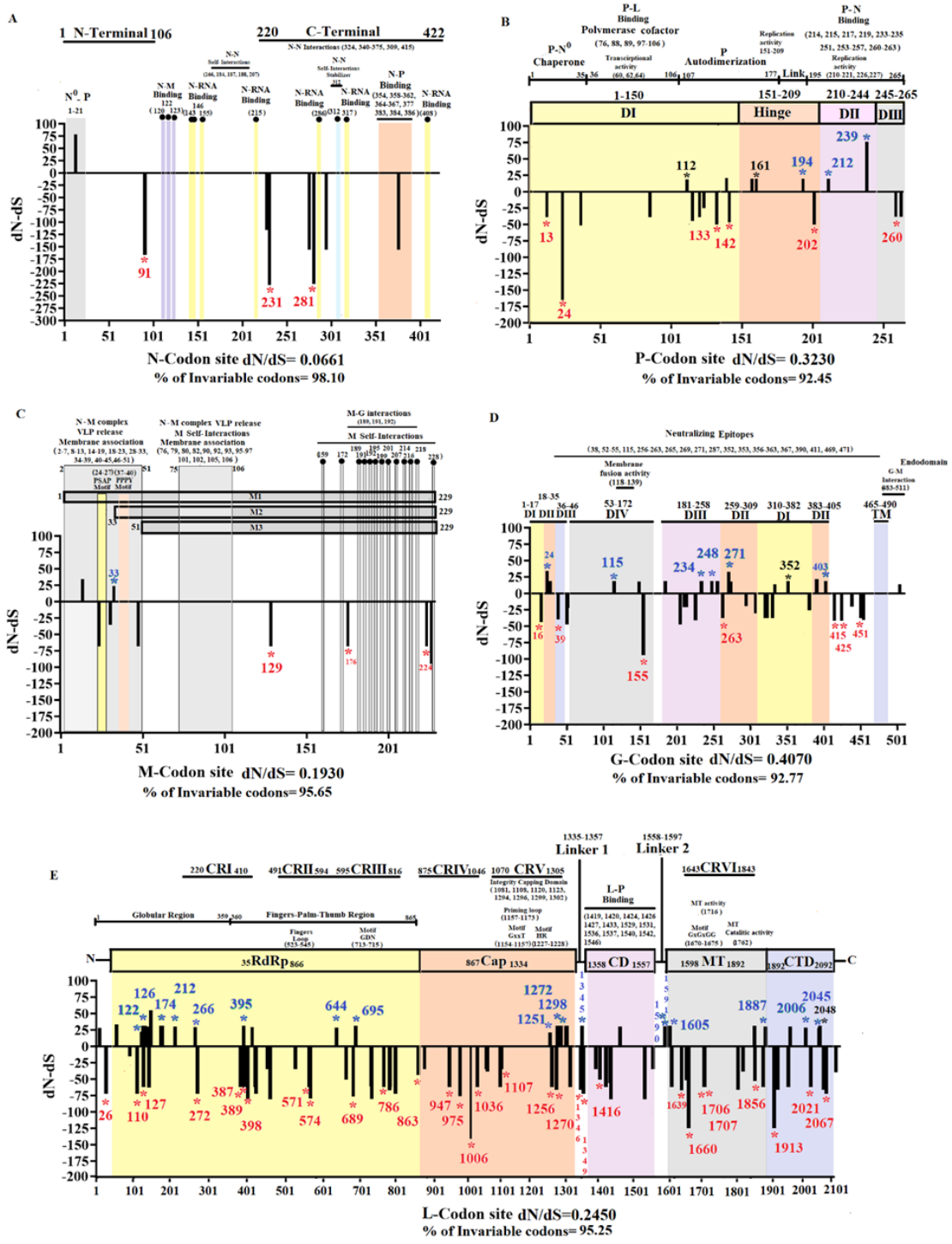


Figure 6. Functional gene evolutionary dynamics of the epidemic VSIV lineage. Graphics represent the dN-dS ratios for specific codon sites at A) Gene N (nucleoprotein), B) Gene P (Phosphoprotein), C) Gene M (Matrix protein), D) Gene G (Glycoprotein), and E) Gene L (Large polymerase). Analyses were conducted using SLAC. Codon sites under positive and purifying selection (identified by MEME and FEL) are highlighted at specific black bars with green and red asterisks, respectively. Similarly, the specific gene location of these codons is indicated by blue and red numbers. Bars highlighted by black asterisks and numbers indicate codon sites identified as relevant by the Metadata-driven comparative analysis but evolving under neutrality based on MEME and FEL analyses. Information about functional sites, relevant motifs, and residues encoded by multiple codon sites at different genes are also indicated. Numbers in parentheses indicate codon positions linked with key residues associated with diverse functions in the viral proteome. The information about functional sites at

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

different viral proteins was obtained from the following publications: Nucleoprotein [36,37,42–44], Phosphoprotein [45–51], Matrix protein [38,40,52,53], Glycoprotein [41,54–59], and Polymerase [60–63].

3.5. Genetic Distance within the Epidemic Lineage Positively Correlates with the Geographical Range of Circulation

Finally, another aim of this study was to obtain further insights into potential factors influencing the evolutionary dynamics of this epidemic lineage. Hence, we attempted to determine the correlation between pairwise genetic and geographical distances. We used geographical information associated with the Colorado dataset. This dataset included isolates representing three out of the four main subpopulations identified in this study.

The analysis results indicated that the spatial distribution of groups 1, 3, and 4 was confined to 24 counties, including 45 cities in Colorado (Figure 7A). The linear regression analysis revealed a significantly positive correlation between genetic and geographical distance values (Figure 7B), suggesting that geographical factors play a significant role in the overall genetic variability of this lineage (Figure 7B).

Once we established a positive correlation between genetic and geographical distances, we attempted to understand if positive selection at multiple isolates might be associated with their circulation in specific geographical settings. We conducted a hierarchical cluster analysis (results were presented as a constellation plot) (Figure 8). Subsequently, based on ANOVA and supported by Tukey's honest significance test (α=0.05), we determined three significant distinct geographical zones of circulation in Colorado (Figure 8). In this context, zone 1 comprised counties and cities associated exclusively with most of the isolates related to the genetic group 1, while zones 2 and 3 included indistinctly isolates from groups 3 and 4. Only four isolates (IN1019COE1, IN10COE5, IN1019COE6, and IN1019COE8) from group 1 circulated in the geographical zone 2. Considering the general geographical pattern distribution of different isolates, we may emphasize the potential role of codon P-239 in the adaptation of isolates associated with the two main phylogenetic clusters to circulate in distinct geographical settings (Figure 8).

All codons detected at leaf nodes under positive selection were linked to isolates indistinctly circulating across all three geographical zones. The only exceptions were codons G-24 and G-271. Interestingly, mutations at codon G-24 were associated with isolates from groups 1 (IN0719COE52) and 3 (IN1019COE2), circulating in two distinct geographical zones (1 and 2). Conversely, codon G-271 was associated with two isolates circulating in geographical zone 2 (Figure 8). Interestingly, as previously stated, the selection of this codon involved single isolates from genetic groups 1 and 4. The close geographical proximity between the places where these isolates were circulating may be possible to suggest the potential adaptive role of G-271.

Finally, our analysis revealed a clear association between isolates selecting codons L-1345 and L-1605 and their circulation at specific geographical settings in geographical zone 2. In the case of codon L-1345, the selection of this codon was associated with the circulation of two isolates (IN1019COE5 and IN1019COE6) from group 1, which were circulating at the county of Garfield and the city of Carbondale (Figure 8). Interestingly, the remaining isolate from group I (IN1019COE1) circulating in the geographical zone 2 at the county of Pitkin and the city of Basalt did not show evidence of selection at codon L-1345. The latter suggests that ecological conditions in counties Garfield and Pitkin may represent two different challenges in terms of adaptation.

Nevertheless, the selection of codon L-1605 was also associated with isolates from genetic group 4 circulating in the contiguous counties of Delta and Mesa (geographical zone 2), indicating the potential role of codon L-1605 in the adaptation of these isolates.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

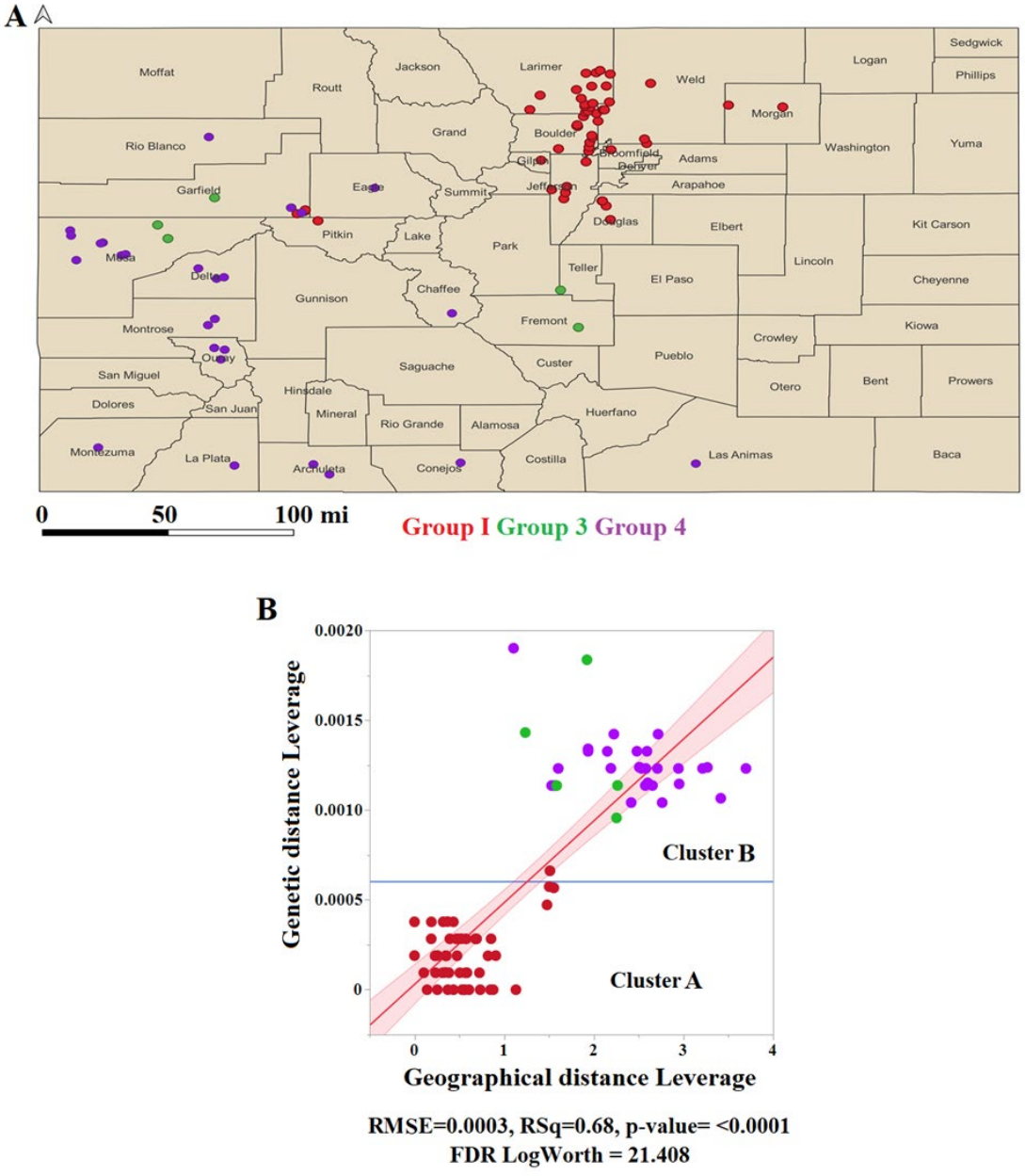
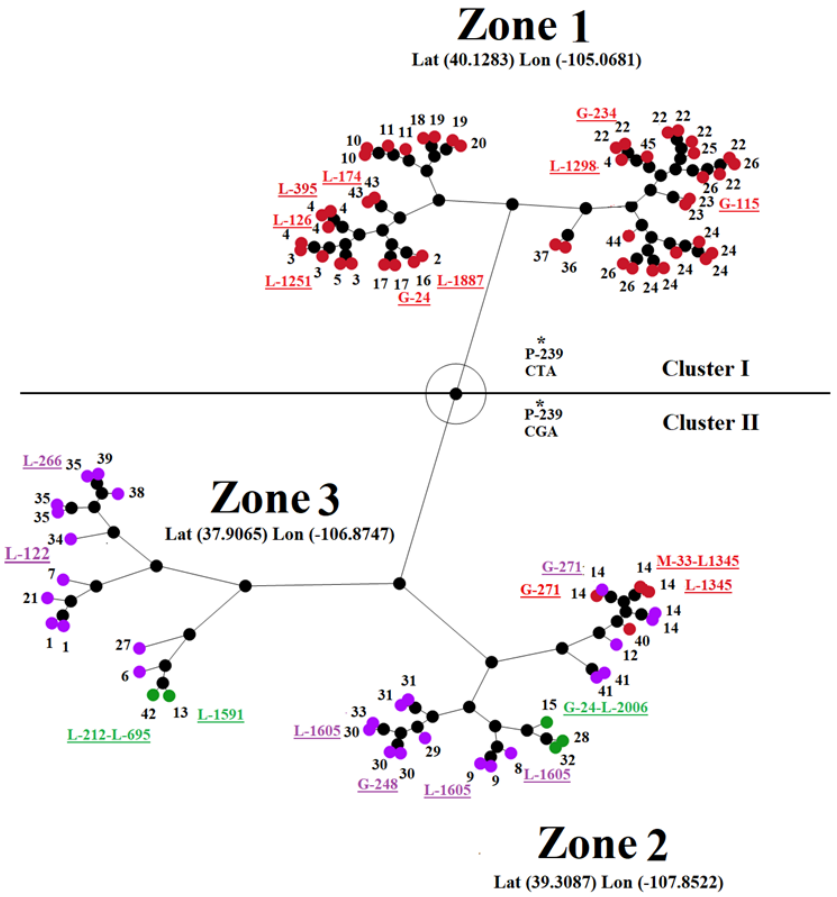


Figure 7. Correlation between genetic and geographical distances using the Colorado dataset as a model. Authors should discuss the results and how they can be interpreted from the perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1



County	City	Code
Archuleta	Pagosa Springs	1
Boulder	Boulder	2
Boulder	Lafayette	3
Boulder	Longmont	4
Broomfield	Broomfield	5
Chaffee	Salida	6
Conejos	La Jara	7
Delta	Cedaredge	8
Delta	Hotchkiss	9
Douglas	Littleton	10
Douglas	Sedalia	11
Eagle	Edwards	12
Fremont	Canon City	13
Garfield	Carbondale	14
Garfield	Rifle	15
Gilpin	Black Hawk	16
Jefferson	Arvada	17
Jefferson	Conifer	18
Jefferson	Evergreen	19
Jefferson	Indian Hills	20
La Plata	Bayfield	21
Larimer	Berthoud	22
Larimer	Estes Park	23
Larimer	Fort Collins	24
Larimer	Johnstown	25
Larimer	Loveland	26
Las Animas	Trinidad	27
Mesa	De Beque	28
Mesa	Glade Park	29
Mesa	Grand Junction	30
Mesa	Loma	31
Mesa	Mesa	32
Mesa	Palisade	33
Montezuma	Cortez	34
Montrose	Montrose	35
Morgan	Ft. Morgan	36
Morgan	Orchard	37
Ouray	Ridgeway	38
Ouray	Ridgway	39
Pitkin	Basalt	40
Rio Blanco	Meeker	41
Teller	Florissant	42
Weld	Brighton	43
Weld	Eaton	44
Weld	Longmont	45

Figure 8. Geographical distribution of isolates displaying codons under positive selection. A constellation plot is shown, depicting the results of a hierarchical cluster analysis based on the latitude and the longitude coordinates where different isolates were collected. Red, green, and purple dots denote isolates belonging to genetic groups 1, 3, and 4, respectively. Different geographical zones determined by ANOVA are indicated. Different codons under positive selection were highlighted next to specific dots to see potential associations between codons at positive selection and their presentation in specific geographical zones. The numbers next to the dots correspond to specific counties and cities in Colorado.

4. Discussion

The emergence of VSV in the US has been a recurrent event during recent decades [4,6–8]. There is solid evidence of the relationship between genetic lineages circulating in Mexico and the ones producing epidemic events in the US [4,6,8]. However, very little is currently known about the role of natural selection in driving the emergence and maintenance of epidemic lineages in the US. Here, based on a representative number of full-length viral sequences, we present for the first time the evolutionary dynamic signatures of an epidemic VSIV lineage that affected the US during 2019-2020.

Consistently with previous studies [4,6,8], our phylogenetic analysis revealed that the closest genetic relative of this epidemic lineage were VSIV isolates circulating in Chiapas – an endemic zone of Mexico. A notable feature of this epidemic lineage is the presence of a nucleotide insertion in the intergenic region between G and L genes, a signature also found in the ancestral isolates from Mexico. This is an interesting finding as insertions in this intergenic region were previously considered a hallmark just for VSIV isolates from a Central America geographical origin [64], indicating that the presence of insertions in intergenic regions is a condition also present in lineages from North America. We currently don't have an explanation for the possible role of this insertion in the epidemic lineage. Our study identified two genotypes distinguished by a 14-nucleotide and a 22-nucleotide insertion.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

Interestingly, all viruses carrying the 22-nucleotide insertion were categorized into the genetic group 3, the less prevalent group during the outbreak. This observation might suggest that the length of insertion may provide a potential adaptive advantage in nature. In vitro, studies have reported the presence of nucleotide insertions in noncoding regions in VSIV genomes in response to evolution on a regimen of constant alternating passages between mammalian and insect cells under laboratory conditions [65], a situation that may be consistent with the evolutionary history an epidemic VSIV lineage in the field.

Although the high degree of conservation of the polyadenylation signal at different intergenic regions in VSV has been reported [32], we found that the length of this signal is not conserved due to the insertion of extra uracil residues in multiple lineages. Interestingly, studies have shown that an increase in the length of the polyadenylation signal from 7 to 14 uracil impaired the ability of the polymerase to initiate downstream mRNA synthesis [32]. Based on this result, we can hypothesize that the observed increase in the uracil tail in a minority of lineages (having 8 and 9 uracil tails) during the outbreak in different intergenic regions might have resulted in a selective disadvantage.

Previous studies have described complex evolutionary dynamics associated with the emergence of arboviral populations, including bottlenecks, founder effects, genetic drift, and natural selection [66,67] (Appendix B). Our results indicated that the epidemic VSIV lineage 2019-2020 diversified into at least four subpopulations during the outbreak. This situation is consistent with previous VSV outbreaks in the US [4,6,8].

We used two different strategies to elucidate the possible role of evolutionary forces in the epidemic dynamics of VSIV lineage 2019-2020. First, we used the Metadata-driven Comparative analysis tool to identify positions in the genome that were significantly different between the four subpopulations. Our findings were consistent with a previous study on VSV showing a positive selection of synonymous mutations during in vitro (selective environments) passaging experiments on VSIV populations [68]. The above can be supported by evidence collected from other viral species [69], suggesting the potential adaptive role of synonymous mutations during the evolution of the epidemic VSIV lineage. Interestingly, the ancestral reconstruction analysis revealed that the fixation of these mutations occurred at internal nodes and remained preserved after multiple transmission cycles in the field, suggesting a pattern driven by natural selection rather than genetic drift. In fact, some of these mutations in previous VSIV lineages circulating in Central and South America emphasize their potential relevance in the adaptation of VSIV in nature. However, we currently lack a clear explanation for the possible functional role of synonymous mutations during the evolution of this epidemic lineage. A previous study on VSIV proposed an interesting hypothesis regarding the potential role of synonymous mutations in immune response evasion in insects produced by RNA interference (RNAi) [68]. Additionally, the role of synonymous mutations in escaping host antiviral RNAi immunity has been experimentally probed during infections with the white spot syndrome virus in shrimp [70]. Considering the different number of vectors identified during the VSIV outbreak in the US in 2020 [71], it is feasible to propose the use of the synonymous mutations identified in our study as markers for future studies to test the role of these mutations in evasion of the immune response by VSV in relevant vector species in the field.

Our second strategy focused on using evolutionary methods that detect sites under positive or purifying selection using dN/dS ratios in a codon-based phylogenetic framework. According to these approaches, we can emphasize the role that purifying selection has in the evolution of VSIV in nature. This result is consistent with previous publications indicating the dominance of purifying selection during the evolution of arboviral populations, a condition that has been associated with the need to replicate in both vertebrate and invertebrate hosts [66,72], or to avoid the high rates of deleterious mutations seen during the evolution of arboviruses [73]. Our analysis revealed that the action of purifying selection was not uniform among different viral genes, strongly suggesting that the functionality of its proteins influences the evolutionary dynamics of VSIV in nature. In this context, the strongest levels of purifying selection were inferred in N and M genes, possibly associated with the critical roles of N and M proteins during the infectious cycle of VSV. The N protein plays a crucial

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

role in protecting the viral genome, in addition to transcription and replication activities [74], while the M protein has a significant function in immune evasion [75], viral assembly and budding [52].

Conversely, we identified P and G genes as the main drivers of the divergence of VSIV in nature. The P protein primarily functions as a polymerase cofactor (linking L-N proteins) for transcription and replication activities and as a chaperone for the proper encapsulation between nascent N proteins and viral RNA [48]. In our study, we identified codon sites linked to specific functional domains where nonsynonymous mutations accumulated during the evolution of the epidemic lineage. Most nonsynonymous mutations were detected in the dimerization domain. In this sense, previous studies showed controversial results about the potential role of this region in the viral growth in cell culture [48,76]. However, none of these mutations were subject to positive selection, suggesting the potentially neutral nature of these changes. Previous amino acid comparisons among different vesiculoviruses indicate that nonsynonymous changes found in the epidemic lineage at codon positions 112, 140, 158, and 151 were associated with variable positions at this functional region [48], supporting our hypothesis about the possible neutral effect of these mutations.

On the other hand, the codons under positive selection at P-194 and P-212 were associated with the hinge and domine II regions of the P protein. Interestingly, experimental deletions in these regions involving codons P-194 and P-212 have adverse effects on the replication of VSIV [47], suggesting a potential phenotypic effect of these mutations at these codons. Additionally, our analyses identified under positive selection codon P-239, located in domain II, associated with the P-N binding region. The interaction between P and N proteins in this region plays an important role in the replication of VSIV [50]. Future studies are needed to identify the possible relevance of P-239 in the infectious cycle of VSIV.

The G protein plays a role in receptor recognition on the host cell surface and triggers membrane fusion after endocytosis of the virion [59]. Our findings indicate that the epidemic lineage accumulated multiple codons under positive selection along different functional domains during the outbreak. Among the codons under positive selection, we can highlight codons G-115 and G-271, both associated with epitopes identified during the evolution of VSIV in the presence of polyclonal antibodies [41]. This result suggests that viruses with mutations at these positions may represent neutralization escape mutant phenotypes. Furthermore, another variable codon, G-352, was identified in association with an epitope during the evolution of the epidemic lineage [77]. As mentioned above, mutation $GCT_{(A)}$-$GTT_{(V)}$ at codon G-352 was linked to the emergence of subpopulation G2, which was dominant during the 2020 outbreak, suggesting that G2 might have represented a lineage capable of escaping neutralization. However, based on the results of MEME and FEL analyses, which categorized G-352 as a neutrally evolving site, we may suggest that fixation of mutation $GCT_{(A)}$-$GTT_{(V)}$ in G2 might have been a result of genetic drift or founder effects rather than natural selection. This situation is expected during the evolution of arboviral populations [66,67]. The relevance of these results will need to be confirmed through future experimental studies.

The L gene harbored most of the sites identified as being under positive selection during the evolution of this epidemic lineage. Similar to the G gene, mutations were found at multiple functional sites. However, the potential effect of various sites under positive selection found in this study will require future studies to confirm their possible phenotypic effect. Numerous codon sites under positive selection are located in conserved regions of the polymerase CRI, CRIII, CRV, and CRVI. A notable finding in our study was the detection of codon L-1784 at CRVI under positive selection, which may be linked to the emergence of the epidemic lineage. The mutation at codon L-1784 is not associated with any critical residue at CRVI, which functions as [ribose-2′-O]-methyltransferase in the VSV polymerase [61]. L-1784 is close to the highly conserved residue K-1795. Mutations at K-1796 have been associated with defects in plaque formation, replication, and mRNA synthesis [61].

Based on ancestral reconstruction analyses, we were able to describe two phylogenetic patterns during the evolution of the epidemic lineage in the US. The most common pattern involving sites under positive selection was associated with the selection of specific codon sites at the leaf nodes of the phylogenetic tree, denoting the potential adaptation of multiple strains at an individual level. However, considering the lack of persistence of these mutations in the population and based on

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

previous publications [78,79], these mutations might have constituted potential evolutionary dead ends, becoming deleterious in the long term. Alternatively, we observed potential advantageous mutations that recently emerged under outbreak conditions. Many of those mutations are expected to fall first toward leaf nodes rather than internal nodes [78]. An excellent example is the codon under positive selection L-1298, which was tracked at leaf nodes in the epizootic lineage but fixed at internal nodes in viruses of Central and South American origin. It was the only codon identified by FEL under positive selection at internal nodes, supporting the potential relevance of L-1298 in producing adaptation at the population level. Additionally, we identified two codons under positive selection (G-24 and G-271) at the leaf nodes, each associated with two iso lates linked to two specific genetic groups, showing the potential adaptive advantage of these mutations. However, due to the limited number of sequences used in our study, these mutations' true prevalence and recurrence may have been underestimated.

The second pattern was associated with an evolutionary scenario previously described in HIV-1 [79], where from the total of codons identified under positive selection (in our study: P-239, L-1354, and L-1605), only a minimal number of mutations at these codons can be expected to be fixed at internal nodes and persist in the population after multiple transmission cycles, suggesting their potential adaptive advantage at the population level. In our study, we identified some nonsynonymous mutations that, despite being tracked at internal nodes and persisting during multiple transmission cycles, no evidence of positive selection was detected by MEME and FEL analyses in the codons associated with these mutations. This was the case for codons P-112, P-161, and G-352, all detected by our metadata-driven comparative analysis. We can offer two possible explanations. First, mutations at P-112, P-161, and G-352 might have arisen in the population due to genetic drift or founder effects [67]. Second, dN/dS approaches may lack power for short evolutionary timescales and cannot detect all types of positive selection [78]. For example, codon G-352, which encodes an amino acid linked with a neutralizing epitope [77], has a plausible mechanism for directional selection (escape). However, as mentioned above, since all this methodology is based on in silico approaches, experimental evidence is needed to understand the potential phenotypic effect of mutations at these codons.

Finally, we found a correlation between geographic and genetic distances among subpopulations that circulated in Colorado. This finding was consistent with a previous study [80], suggesting that ecological factors, rather than temporal ones, play a dominant role in the evolution of VSV in the field. Our analysis supports the potential relevance of some codon sites under positive selection with effect at a population level, like P-239, clearly associated with the circulation of genetic groups G1 and G3-G4 in distant locations. Other important codons, L-1345 and L-1605, potentially favor the adaptation within groups 1 and 4, respectively. Considering that all viral lineages from Colorado were recovered from horses, it is possible to hypothesize that the possible adaptation produced by P-239, L-1345, and L-1605 might have been linked to the need for VSIV to replicate in distinct insect vector populations. This hypothesis finds support in previous studies that reported the presence of VSIV in multiple insect vectors during the epizootic outbreak in 2020 [71,81]. However, it is important to note that the identification of the different geographical zones in Colorado is based solely on significant differences in geographical distances among them; thus, further studies are also needed to better characterize their ecological differences. Future in vivo studies are needed to comprehend the relevance of these codon positions in the adaptation of VSIV to replicate in distinct insect vectors.

Conversely, several codon sites under positive selection associated with individual adaptation appeared randomly distributed along different geographic zones. This suggests that these sites played a limited role of these sites in the adaptation of these lineages during the outbreak. However, as mentioned earlier, it is crucial to consider that these mutations' true prevalence and recurrence may have been underestimated.

5. Conclusion

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

In summary, our study revealed the complex dynamics involving the evolution of a highly conserved epidemic lineage of VSIV in the US. It highlights the importance of natural selection in the adaptation of VSIV in the field, suggesting that a minimal number of changes in the P, G, and L genes at specific functional regions might be responsible for the adaptation of this lineage during the outbreak. Our study indicates the potential significance of the synonymous mutations during the circulation of this lineage in the US. Additionally, we emphasize the importance of using a combined methodology based on dN/dS and detecting mutations that are increasing their frequencies in the population to identify potential key codon positions that are driving the evolution of VSV lineages during epizootic outbreaks. We consider that the methodology used in this study may serve as a framework for conducting future evolutionary studies on VSV in nature.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org. **Figure S1:** Ancestral reconstruction analysis of the SNPs identified by Metadata-driven comparative analysis; **Figure S2:** Codons under purifying selection predicted by FEL analysis, **Figure S3:** Ancestral reconstruction analysis of mutations at codons under positive selection identified by MEME analysis; **Figure S4:** Ancestral reconstruction analysis at codons L-1298 and L-1784.

Author Contributions: Conceptualization, SZ, SKP and LV-S; methodology, SZ, SKP and LV-S; software, SZ and SKP.; formal analysis, SZ, SKP and LV-S ; investigation, SZ, MB, CR, KR, AP-M, NG-R, LR, CM, CHM, SKP, and LV-S; resources, LR, and CHM; data curation, SZ, MB, CR, KR, AP-M, NG-R, CM, SKP, and LV-S; ; writing—original draft preparation, SZ, SKP and LV-S ; writing—review and editing, SZ, MB, CR, KR, AP-M, NG-R, LR, CM, CHM, SKP, and LV-S ; visualization, SZ, MB, CR, KR, AP-M, NG-R, LR, CM, CHM, SKP, and LV-S ; supervision, LV-S; project administration, LV-S and CHM ; funding acquisition, LR, and CHM All authors have read and agreed to the published version of the manuscript.

Funding: This work was conducted under the USDA Research Service CRIS Project "Predicting and Mitigating Vesicular Stomatitis Virus (VSV) in North America" No. 3022-32000-062-000-D.

Institutional Review Board Statement: Not applicable. This study did not involve human or animal subjects.

Data Availability Statement: All sequences used for this study are available at GenBank database. The alignments of each gene used to conduct the evolutionary analyses in this study are available at https://github.com/spond/pubs/tree/master/VSIV.

Acknowledgments: The authors would like to acknowledge the state and federal animal health officials, private veterinarians who responded to the 2019 VS outbreak in Colorado, and the Colorado State University Diagnostic Laboratory, which was activated as part of NAHLN to test equine samples during the outbreak.

Conflicts of Interest: The authors declare that the research was conducted without any commercial or financial relationships that could potentially create a conflict of interest.

Appendix A

Figure S1: The figure shows the ancestral reconstruction analysis conducted at different SNPs identified by the Metadata-driven comparative analysis, showing the ancestral state at different internal and leaf nodes of each genetic group (G1 to G4). Next to each tree (Left side), is included the information regarding the specific gene-codon where this SNP was identified, nucleotide position at that specific gene, codon change (capital letter represents the position where SNPs were found at specific codons), and information about this specific mutation other VSIV not related to the VSIV 2019-2020 outbreak in the US. Red asterisks indicate the specific internal nodes in the tree where mutations at specific codons were tracked in the ancestral reconstruction analysis.

Figure S2: The figure shows specific codon sites at N, P, M, G and L genes predicted under purifying selection by FEL analysis. This analysis was conducted using the alignment available at https://github.com/spond/pubs/tree/master/VSIV. Sites under purifying selection may represent potential functional sites at different proteins, preserved at VSIV during its evolution in nature.

Figure S3: The figure shows the ancestral reconstruction analysis conducted at different sites involving codons identified under positive selection by MEME analysis in the VSIV epidemic lineage. This analysis was conducted using the alignment available at https://github.com/spond/pubs/tree/master/VSIV. Different genetic groups identified in this study

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

G1 to G4 are highlighted in the tree by color branches in red, blue, green and purpure respectively. Next to each tree (Left side), is included the information regarding the specific gene-codon were identified under positive selection. Black asterisks indicate the specific internal or leaf nodes in the tree where mutations at specific codons under positive were tracked in the ancestral reconstruction analysis. Overall, all mutations at codons under positive selection were tracked in the leaf nodes, the only exceptions were codons P-239, L1345, and L-1605, where mutations were also tracked at internal nodes.

Figure S4: The figure shows the ancestral reconstruction analysis conducted on SLAC at codons L-1298 and L-1784. L-1298 was the only codon identified under positive selection at internal nodes by FEL analysis, indicating its potential relevance in the adaptation of VSIV at population level. During the VSIV outbreak in the US in 2019-2020, we detected L-1298 under positive selection at single leaf node (IN0719COE12). Interestingly, L-1298 appeared under positive selection in VSIV populations of central and south America. The selection of codon $ACC_{(T)}$ was tracked at internal nodes (that represent the sequences of the predicted ancestral VSIV circulating during the outbreak), indicating its potential relevance in the adaptation of VSIV at population level. On the other hand, the detection of positive section of codon 1784, should be considered a relevant fining since mutation $CGG_{(R)}$-$CAG_{(Q)}$ is highly conserved in all viruses associated with the epidemic lineage, stressing its possible relevance in the emergence of the epidemic VSIV lineage in the US.

Appendix B (Definitions Adapted to This Study)

Positive selection or episodic diversifying selection: Natural selective force that promotes adaptation and innovation by increasing amino acid diversity in a viral protein.

Purifying or negative selection: Natural selective force that favors the amino acid conservation in a viral protein. It results in a stabilizing selection that preserve the functionality of a protein by removing deleterious alleles from the viral population.

Genetic drift: Random mutations that occur in the viral genome due to chance, random sampling. Genetic drift reduces the efficiency of selection and genetic diversity.

Population bottlenecks: It represents a form of genetic drift which produce an important reduction in the population size of organisms, resulting the fixation of mutations at random, and loss in genetic diversity. Replication in insect vectors represent an important source of population bottlenecks for arboviruses.

Founder effects: It is a form of genetic drift, that results in the fixation of random mutations. It leads to the loss of genetic variability in viral population. In arboviruses it may happen when a susceptible host is infected with a specific viral quasipecies derived from a bottleneck after replication in an insect vector. Then this host moves into a different geographical isolated area, producing the infection of new susceptible hosts, with the consequently dominance of this viral quasipecies this area.

References

1. Velazquez-Salinas, L., Zarate, S., Eschbaumer, M., Pereira Lobo, F., Gladue, D.P., Arzt, J., Novella, I.S., Rodriguez, L.L., 2016. Selective Factors Associated with the Evolution of Codon Usage in Natural Populations of Arboviruses. PLoS One 11, e0159943.
2. Dietzgen, R.G., 2012. Morphology. Genome Organization, Transcription and Replication of Rhabdoviruses, in: Kuzmin, R.G.D.a.I.V. (Ed.), Rhabdoviruses, pp. 5-11.
3. Rodriguez, L.L., 2002. Emergence and re-emergence of vesicular stomatitis in the United States. Virus Res 85, 211-219.
4. Velazquez-Salinas, L., Pauszek, S.J., Zarate, S., Basurto-Alcantara, F.J., Verdugo-Rodriguez, A., Perez, A.M., Rodriguez, L.L., 2014. Phylogeographic characteristics of vesicular stomatitis New Jersey viruses circulating in Mexico from 2005 to 2011 and their relationship to epidemics in the United States. Virology 449, 17-24
5. Pauszek, S.J., Rodriguez, L.L., 2012. Full-length genome analysis of vesicular stomatitis New Jersey virus strains representing the phylogenetic and geographic diversity of the virus. Arch Virol 157, 2247-2251.
6. Rodriguez, L.L., Bunch, T.A., Fraire, M., Llewellyn, Z.N., 2000. Re-emergence of vesicular stomatitis in the western United States is associated with distinct viral genetic lineages. Virology 271, 171-181.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

7. Pelzel-McCluskey, A., Christensen, B., Humphreys, J., Bertram, M., Keener, R., Ewing, R., Cohnstaedt, L.W., Tell, R., Peters, D.P.C., Rodriguez, L., 2021. Review of Vesicular Stomatitis in the United States with Focus on 2019 and 2020 Outbreaks. Pathogens 10.
8. Rainwater-Lovett, K., Pauszek, S.J., Kelley, W.N., Rodriguez, L.L., 2007. Molecular epidemiology of vesicular stomatitis New Jersey virus from the 2004-2005 US outbreak indicates a common origin with Mexican strains. J Gen Virol 88, 2042-2051
9. Hole, K., Nfon, C., Rodriguez, L.L., Velazquez-Salinas, L., 2021. A Multiplex Real-Time Reverse Transcription Polymerase Chain Reaction Assay With Enhanced Capacity to Detect Vesicular Stomatitis Viral Lineages of Central American Origin. Front Vet Sci 8, 783198
10. Goodger, W.J., Thurmond, M., Nehay, J., Mitchell, J., Smith, P., 1985. Economic impact of an epizootic of bovine vesicular stomatitis in California. J Am Vet Med Assoc 186, 370-373
11. Hayek, A.M., McCluskey, B.J., Chavez, G.T., Salman, M.D., 1998. Financial impact of the 1995 outbreak of vesicular stomatitis on 16 beef ranches in Colorado. J Am Vet Med Assoc 212, 820-823.
12. Velazquez-Salinas, L., Pauszek, S.J., Stenfeldt, C., O'Hearn, E.S., Pacheco, J.M., Borca, M.V., Verdugo-Rodriguez, A., Arzt, J., Rodriguez, L.L., 2018. Increased Virulence of an Epidemic Strain of Vesicular Stomatitis Virus Is Associated With Interference of the Innate Response in Pigs. Front Microbiol 9, 1891.
13. Rozo-Lopez, P., Pauszek, S.J., Velazquez-Salinas, L., Rodriguez, L.L., Park, Y., Drolet, B.S., 2022. Comparison of Endemic and Epidemic Vesicular Stomatitis Virus Lineages in Culicoides sonorensis Midges. Viruses 14.
14. Bertram, M.R., Rodgers, C., Reed, K., Velazquez-Salinas, L., Pelzel-McCluskey, A., Mayo, C., Rodriguez, L., 2023. Vesicular stomatitis Indiana virus near-full-length genome sequences reveal low genetic diversity during the 2019 outbreak in Colorado, USA. Front Vet Sci 10, 1110483
15. Hole, K., Buchanan, C., Lung, O., Babiuk, S., Nfon, C., Navarro-Lopez, R., Gomez-Romero, N., Rodriguez, L.L, Bertram, R.M., Mire, C., and Velazquez-Salinas, L. Near-full length Vesicular Stomatitis Indiana Virus Genome Sequences Representative of Endemic Strains Circulating in Mexico Genome Announc.
16. Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22, 4673-4680.
17. Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 35, 1547-1549.
18. Pickett, B.E., Liu, M., Sadat, E.L., Squires, R.B., Noronha, J.M., He, S., Jen, W., Zaremba, S., Gu, Z., Zhou, L., Larsen, C.N., Bosch, I., Gehrke, L., McGee, M., Klem, E.B., Scheuermann, R.H., 2013. Metadata-driven comparative analysis tool for sequences (meta-CATS): an automated process for identifying significant sequence variations that correlate with virus attributes. Virology 447, 45-51.
19. Hudson, R.R., Slatkin, M., Maddison, W.P., 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132, 583-589.
20. Pond, S.L., Frost, S.D., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676-679
21. Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20, 289-290.
22. Paradis, E., 2010. pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 26, 419-420.
23. Velazquez-Salinas, L., Zarate, S., Eberl, S., Gladue, D.P., Novella, I., Borca, M.V., 2020. Positive Selection of ORF1ab, ORF3a, and ORF8 Genes Drives the Early Evolutionary Trends of SARS-CoV-2 During the 2020 COVID-19 Pandemic. Frontiers in Microbiology 11.
24. Kosakovsky Pond, S.L., Frost, S.D., 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22, 1208-1222.
25. Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K., Kosakovsky Pond, S.L., 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet 8, e1002764.
26. Weaver, S., Shank, S.D., Spielman, S.J., Li, M., Muse, S.V., Kosakovsky Pond, S.L., 2018. Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. Mol Biol Evol 35, 773-777.
27. Spielman, S.J., Weaver, S., Shank, S.D., Magalis, B.R., Li, M., Kosakovsky Pond, S.L., 2019. Evolution of Viral Genomes: Interplay Between Selection, Recombination, and Other Forces. Methods Mol Biol 1910, 427-468.
28. Wertheim, J.O., Murrell, B., Smith, M.D., Kosakovsky Pond, S.L., Scheffler, K., 2015. RELAX: detecting relaxed selection in a phylogenetic framework. Mol Biol Evol 32, 820-832.
29. Kosakovsky Pond, S.L., Posada, D., Gravenor, M.B., Woelk, C.H., Frost, S.D., 2006. GARD: a genetic algorithm for recombination detection. Bioinformatics 22, 3096-3098.
30. Hipp, A.L., Hall, J.C., Sytsma, K.J., 2004. Congruence versus phylogenetic accuracy: revisiting the incongruence length difference test. Syst Biol 53, 81-89.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

31. Kosakovsky Pond, S.L., Poon, A.F.Y., Velazquez, R., Weaver, S., Hepler, N.L., Murrell, B., Shank, S.D., Magalis, B.R., Bouvier, D., Nekrutenko, A., Wisotsky, S., Spielman, S.J., Frost, S.D.W., Muse, S.V., 2020. HyPhy 2.5-A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. Mol Biol Evol 37, 295-299
32. Barr, J.N., Whelan, S.P., Wertz, G.W., 1997. cis-Acting signals involved in termination of vesicular stomatitis virus mRNA synthesis include the conserved AUAC and the U7 signal for polyadenylation. J Virol 71, 8718-8725.
33. Benndorf, R., Velazquez, R., Zehr, J.D., Pond, S.L.K., Martin, J.L., Lucaci, A.G., 2022. Human HspB1, HspB3, HspB5 and HspB8: Shaping these disease factors during vertebrate evolution. Cell Stress Chaperones 27, 309-323.
34. Lucaci, A.G., Zehr, J.D., Enard, D., Thornton, J.W., Kosakovsky Pond, S.L., 2023. Evolutionary Shortcuts via Multinucleotide Substitutions and Their Impact on Natural Selection Analyses. Mol Biol Evol 40.
35. Zehr, J.D., Pond, S.L.K., Martin, D.P., Ceres, K., Whittaker, G.R., Millet, J.K., Goodman, L.B., Stanhope, M.J., 2022. Recent Zoonotic Spillover and Tropism Shift of a Canine Coronavirus Is Associated with Relaxed Selection and Putative Loss of Function in NTD Subdomain of Spike Protein. Viruses 14.
36. Leyrat, C., Yabukarski, F., Tarbouriech, N., Ribeiro, E.A., Jr., Jensen, M.R., Blackledge, M., Ruigrok, R.W., Jamin, M., 2011. Structure of the vesicular stomatitis virus N(0)-P complex. PLoS Pathog 7, e1002248
37. Green, T.J., Luo, M., 2009. Structure of the vesicular stomatitis virus nucleocapsid in complex with the nucleocapsid-binding domain of the small polymerase cofactor, P. Proc Natl Acad Sci U S A 106, 11713-11718.
38. Dancho, B., McKenzie, M.O., Connor, J.H., Lyles, D.S., 2009. Vesicular stomatitis virus matrix protein mutations that affect association with host membranes and viral nucleocapsids. J Biol Chem 284, 4500-4509
39. Jayakar, H.R., Whitt, M.A., 2002. Identification of two additional translation products from the matrix (M) gene that contribute to vesicular stomatitis virus cytopathology. J Virol 76, 8011-8018.
40. Redondo, N., Madan, V., Alvarez, E., Carrasco, L., 2015. Impact of Vesicular Stomatitis Virus M Proteins on Different Cellular Functions. PLoS One 10, e0131137.
41. Vandepol, S.B., Lefrancois, L., Holland, J.J., 1986. Sequences of the major antibody binding epitopes of the Indiana serotype of vesicular stomatitis virus. Virology 148, 312-325.
42. Green, T.J., Zhang, X., Wertz, G.W., Luo, M., 2006. Structure of the vesicular stomatitis virus nucleoprotein-RNA complex. Science 313, 357-360.
43. Hanke, L., Schmidt, F.I., Knockenhauer, K.E., Morin, B., Whelan, S.P., Schwartz, T.U., Ploegh, H.L., 2017. Vesicular stomatitis virus N protein-specific single-domain antibody fragments inhibit replication. EMBO Rep 18, 1027-1037.
44. Zhou, K., Si, Z., Ge, P., Tsao, J., Luo, M., Zhou, Z.H., 2022. Atomic model of vesicular stomatitis virus and mechanism of assembly. Nat Commun 13, 5980.
45. Chen, M., Ogino, T., Banerjee, A.K., 2006. Mapping and functional role of the self-association domain of vesicular stomatitis virus phosphoprotein. J Virol 80, 9511-9518.
46. Das, S.C., Pattnaik, A.K., 2004. Phosphorylation of vesicular stomatitis virus phosphoprotein P is indispensable for virus growth. J Virol 78, 6420-6430.
47. Das, S.C., Pattnaik, A.K., 2005. Role of the hypervariable hinge region of phosphoprotein P of vesicular stomatitis virus in viral RNA synthesis and assembly of infectious virus particles. J Virol 79, 8101-8112.
48. Gerard, F.C.A., Jamin, M., Blackledge, M., Blondel, D., Bourhis, J.M., 2020. Vesicular Stomatitis Virus Phosphoprotein Dimerization Domain Is Dispensable for Virus Growth. J Virol 94.
49. Gould, J.R., Qiu, S., Shang, Q., Ogino, T., Prevelige, P.E., Jr., Petit, C.M., Green, T.J., 2020. The Connector Domain of Vesicular Stomatitis Virus Large Protein Interacts with the Viral Phosphoprotein. J Virol 94.
50. Green, T.J., Macpherson, S., Qiu, S., Lebowitz, J., Wertz, G.W., Luo, M., 2000. Study of the assembly of vesicular stomatitis virus N protein: role of the P protein. J Virol 74, 9515-9524.
51. Hwang, L.N., Englund, N., Das, T., Banerjee, A.K., Pattnaik, A.K., 1999. Optimal replication activity of vesicular stomatitis virus RNA polymerase requires phosphorylation of a residue(s) at carboxy-terminal domain II of its accessory subunit, phosphoprotein P. J Virol 73, 5613-5620.
52. Gaudier, M., Gaudin, Y., Knossow, M., 2002. Crystal structure of vesicular stomatitis virus matrix protein. EMBO J 21, 2886-2892.
53. Lichty, B.D., McBride, H., Hanson, S., Bell, J.C., 2006. Matrix protein of Vesicular stomatitis virus harbours a cryptic mitochondrial-targeting motif. J Gen Virol 87, 3379-3384.
54. Keil, W., Wagner, R.R., 1989. Epitope mapping by deletion mutants and chimeras of two vesicular stomatitis virus glycoprotein genes expressed by a vaccinia virus vector. Virology 170, 392-407.
55. Munis, A.M., Tijani, M., Hassall, M., Mattiuzzo, G., Collins, M.K., Takeuchi, Y., 2018. Characterization of Antibody Interactions with the G Protein of Vesicular Stomatitis Virus Indiana Strain and Other Vesiculovirus G Proteins. J Virol 92.
56. Nikolic, J., Belot, L., Raux, H., Legrand, P., Gaudin, Y., A, A.A., 2018. Structural basis for the recognition of LDL-receptor family members by VSV glycoprotein. Nat Commun 9, 1029.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

57. Roche, S., Bressanelli, S., Rey, F.A., Gaudin, Y., 2006. Crystal structure of the low-pH form of the vesicular stomatitis virus glycoprotein G. Science 313, 187-191.
58. Roche, S., Rey, F.A., Gaudin, Y., Bressanelli, S., 2007. Structure of the prefusion form of the vesicular stomatitis virus glycoprotein G. Science 315, 843-848.
59. Roche, S., Albertini, A.A., Lepault, J., Bressanelli, S., Gaudin, Y., 2008. Structures of vesicular stomatitis virus glycoprotein: membrane fusion revisited. Cell Mol Life Sci 65, 1716-1728.
60. Galloway, S.E., Wertz, G.W., 2008. S-adenosyl homocysteine-induced hyperpolyadenylation of vesicular stomatitis virus mRNA requires the methyltransferase activity of L protein. J Virol 82, 12280-12290.
61. Li, J., Fontaine-Rodriguez, E.C., Whelan, S.P., 2005. Amino acid residues within conserved domain VI of the vesicular stomatitis virus large polymerase protein essential for mRNA cap methyltransferase activity. J Virol 79, 13373-13384.
62. Liang, B., Li, Z., Jenni, S., Rahmeh, A.A., Morin, B.M., Grant, T., Grigorieff, N., Harrison, S.C., Whelan, S.P.J., 2015. Structure of the L Protein of Vesicular Stomatitis Virus from Electron Cryomicroscopy. Cell 162, 314-327.
63. Ruedas, J.B., Perrault, J., 2014. Putative domain-domain interactions in the vesicular stomatitis virus L polymerase protein appendage region. J Virol 88, 14458-14466.
64. Rodriguez, L.L., Pauszek, S.J., Bunch, T.A., Schumann, K.R., 2002. Full-length genome analysis of natural isolates of vesicular stomatitis virus (Indiana 1 serotype) from North, Central and South America. J Gen Virol 83, 2475-2483.
65. Novella, I.S., Ebendick-Corpus, B.E., Zarate, S., Miller, E.L., 2007. Emergence of mammalian cell-adapted vesicular stomatitis virus from persistent infections of insect vector cells. J Virol 81, 6664-6668.
66. Weaver, S.C., 2006. Evolutionary influences in arboviral disease. Curr Top Microbiol Immunol 299, 285-314.
67. Weaver, S.C., Forrester, N.L., Liu, J., Vasilakis, N., 2021. Population bottlenecks and founder effects: implications for mosquito-borne arboviral emergence. Nat Rev Microbiol 19, 184-195.
68. Novella, I.S., Zarate, S., Metzgar, D., Ebendick-Corpus, B.E., 2004. Positive selection of synonymous mutations in vesicular stomatitis virus. J Mol Biol 342, 1415-1421.
69. Bailey, S.F., Alonso Morales, L.A., Kassen, R., 2021. Effects of Synonymous Mutations beyond Codon Bias: The Evidence for Adaptive Synonymous Substitutions from Microbial Evolution Experiments. Genome Biol Evol 13.
70. Sun, Y., Zhang, Y., Zhang, X., 2019. Synonymous SNPs of viral genes facilitate virus to escape host antiviral RNAi immunity. RNA Biol 16, 1697-1710.
71. McGregor, B.L., Rozo-Lopez, P., Davis, T.M., Drolet, B.S., 2021. Detection of Vesicular Stomatitis Virus Indiana from Insects Collected during the 2020 Outbreak in Kansas, USA. Pathogens 10.
72. Jerzak, G., Bernard, K.A., Kramer, L.D., Ebel, G.D., 2005. Genetic variation in West Nile virus from naturally infected mosquitoes and birds suggests quasispecies structure and strong purifying selection. J Gen Virol 86, 2175-2183.
73. Holmes, E.C., 2003. Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. J Virol 77, 11296-11298.
74. Patil, G., Xu, L., Wu, Y., Song, K., Hao, W., Hua, F., Wang, L., Li, S., 2020. TRIM41-Mediated Ubiquitination of Nucleoprotein Limits Vesicular Stomatitis Virus Infection. Viruses 12.
75. Marquis, K.A., Becker, R.L., Weiss, A.N., Morris, M.C., Ferran, M.C., 2020. The VSV matrix protein inhibits NF-kappaB and the interferon response independently in mouse L929 cells. Virology 548, 117-123.
76. Bloyet L, Morin BBrusic V, Gardner E, Ross RA, Vadakkan T, Kirchhausen T, Whelan SPJ 2020. Oligomerization of the Vesicular Stomatitis Virus Phosphoprotein Is Dispensable for mRNA Synthesis but Facilitates RNA Replication. J Virol 94:10.1128/jvi.00115-20.
77. Munis, A.M., Tijani, M., Hassall, M., Mattiuzzo, G., Collins, M.K., Takeuchi, Y., 2018. Characterization of Antibody Interactions with the G Protein of Vesicular Stomatitis Virus Indiana Strain and Other Vesiculovirus G Proteins. J Virol 92.
78. Geoghegan, J.L., Holmes, E.C., 2018. The phylogenomics of evolving virus virulence. Nat Rev Genet 19, 756-769.
79. Pond, S.L., Frost, S.D., Grossman, Z., Gravenor, M.B., Richman, D.D., Brown, A.J., 2006. Adaptation to different human populations by HIV-1 revealed by codon-based analyses. PLoS Comput Biol 2, e62.
80. Rodriguez, L.L., Fitch, W.M., Nichol, S.T., 1996. Ecological factors rather than temporal factors dominate the evolution of vesicular stomatitis virus. Proc Natl Acad Sci U S A 93, 13030-13035.
81. Young, K.I., Valdez, F., Vaquera, C., Campos, C., Zhou, L., Vessels, H.K., Moulton, J.K., Drolet, B.S., Rozo-Lopez, P., Pelzel-McCluskey, A.M., Peters, D.C., Rodriguez, L.L., Hanley, K.A., 2021. Surveillance along the Rio Grande during the 2020 Vesicular Stomatitis Outbreak Reveals Spatio-Temporal Dynamics of and Viral RNA Detection in Black Flies. Pathogens 10.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2024 doi:10.20944/preprints202410.0585.v1

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.