Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

GENETIC DIVERSITY, PHYLOGENETIC RELATIONSHIPS AND RECOMBINATION SIGNAL IN THE COAT PROTEIN GENE OF CITRUS TRISTEZA VIRUS ISOLATES FROM PAKISTAN

Kalsoom Bibi

Attaurehman School of Applied Science, Islamabad

kalsoom0435@gmail.com

DOI: https://doi.org/10.5281/zenodo.22272191

Keywords
Citrus tristeza virus, Genetic diversity, Coat protein gene, Phylogenetic analysis, Recombination detection, Pakistani isolates, MAFFT, PhyML, GARD, Citrus virology

Article History
Received: 18 January 2026
Accepted: 03 March 2026
Published: 17 March 2026

Corresponding Author: *
Kalsoom Bibi

Abstract

Citrus tristeza virus (CTV Closterovirus genus, Closteroviridae) is a virus that is one of the most economic diseases of citrus worldwide. In this study, the coat protein (CP) gene of ten Pakistani CTV field isolates (Haripur, Fatehjang, Islamabad ×2, Nowshera, Sawabi, Faislabad, Sargodha and Khanpur) was retrieved from GenBank and analysed alongside a comparative reference panel comprising two additional CTV CP sequences, five CTV ORF1a genotype-marker fragments (genotypes B3 and D1–D4). Five citrus host EST cDNA sequences and representative sequences of three other citrus infecting viruses (Citrus leaf blotch virus, Citrus bright spot virus and Citrus leprosis virus N) for a total of 31 raw GenBank records. Sequences were aligned with MAFFT trimmed to a 456 nucleotide informative block using BMGE and the resulting alignment was used to build a max likelihood tree with PhyML. A recombination detection algorithm based on a genetic algorithm (GARD) suggested a single recombination breakpoint at position 259 of the alignment with a low AICc score of 1197.25 compared to a single tree model and 696 candidate topologies screened. It meaning the different parts of the analysed sequence have evolved differently. The pairwise identity between the ten Pakistani isolates was observed between 90.8% to 99.8% and there was no indication of a single tight monophyletic cluster in the ML tree suggesting co circulation of divergent CTV genotypes in Pakistan.

1 Introduction

Citrus tristeza virus (CTV) is a phloem-limited, single-stranded positive-sense RNA virus with one of the largest genomes of all plant viruses (approx. 19.3 kb) which are divided into over a dozen open reading frames (ORFs). CTV isolates are usually clustered into six different genotypes based on sequence divergence T3/T30, T36, VT, RB, B165/B2 and the D and B subgroups as used in some genotyping schemes these isolates vary in geographic distribution, vector transmissibility and symptom severity ranging from mild to devastating forms of quick decline and stem pitting. Coat protein and ORF1a molecular surveys are routinely conducted in a citrus growing region to determine the presence of which CTV genotypes and to identify potential inter genotype recombination that may produce new biological phenotypes.

A set of 10 CTV coat-protein gene sequences obtained from Pakistani citrus isolates and a larger comparative set obtained in the same search from Gen Bank were downloaded for this analysis which included additional CTV CP and ORF1a reference sequences host citrus ES Ts and three unrelated viruses that infect citrus. Here we use this heterogeneous panel as it is intended for bio informatic use the non-Pakistani (CTV) sequences serve as genotype context for the Pakistani sequences and non-CTV sequences (host ESTs, CLBV, Ci BSV, Ci LV-N) are used as divergence controls out group material for the alignment, tree and

recombination pipeline. This framing is the same as that of the raw sequence alignment, tree and GARD files that were uploaded and is explicitly stated so that the composition of the dataset is clear to readers. The most common CTV marker that has been sequenced broadly is coat protein (p25 CP) gene and has been used to characterise the diversity of isolates from many citrus growing countries such as repeated surveys in Pakistan (Khan et al. 2022; Shafique et al. 2024a, 2024b). Data from the surveys conducted in Sargodha district one of the major citrus growing areas of the country reported up to 90% incidence of disease in Mosambi 74% in Feutrell s early and 57% in Kinnow with maximum severity in Kot Momin tehsil (Khan et al. 2022).

2Materials and Methods

A total of 31 nucleotide sequences were obtained from GenBank (Table 1). Ten complete CTV coat-protein and SY553) with five CTV ORF1a fragments corresponding to genotype B3 and D1-D4 gene sequences (672 bp) from Pakistani isolates and two additional CTV CP sequences (isolates SY550 (404 bp) were included as genotype reference material. The panel was completed with five Citrus aurantium (sour orange) EST cDNA clones from a Citrus Tristeza Virus-challenged bark library two Citrus leaf blotch virus CP sequences one Citrus bright spot virus nucleocapsid sequence and four Citrus leprosis virus N glycoprotein (ORF5) sequences. Two records in the raw download were exact duplicate accessions (EU918609.1 and DT214395.1) and were counted as replicates not independent isolates in downstream counts.

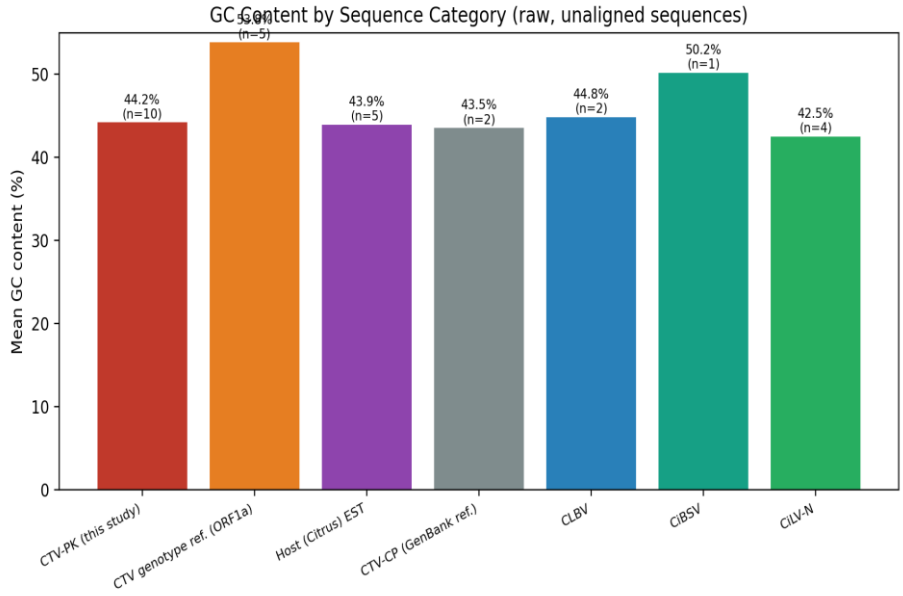

Figure 1. Mean GC content (%) by sequence category in the raw (unaligned) dataset. Bars show category means with sample size (n); the ten Pakistani CTV-CP isolates cluster near 44% GC, similar to the CTV CP reference and host EST sequences, while the ORF1a genotype markers (which include long stretches of a different more GC-rich genome region) and CiBSV are higher.

Alignment and trimming

MAFFT was used for all 31 sequences to align them. Multiple sequence alignment was trimmed with BMGE (Block Mapping and Gathering with Entropy) for poor ambiguous alignment resulting in a curated alignment block of 456nt that was used for all subsequent phylogenetic and recombination analysis (29 unique sequence names after deduplication 27 sequences retained in the GARD run after further quality filtering). The sequences start with an ATG codon and terminate in a stop codon in frame with the expected CP reading frame for CTV with each sequence interrupted by a 224 residue coat protein beginning with MDDETKKLKN... and ending with ...RQLGKFNTR both a single residue variant of the original.

Phylogenetic reconstruction

The BMGE trimmed alignment was used to reconstruct a maximum likelihood phylogeny by PhyML. A visualisation of the Newick tree was coloured by sequence category (Pakistani CTV CP isolates, other CTV CP isolates, host ESTs) and the three other citrus viruses to assist interpretation.

Recombination screening (GARD)

A 456 bp alignment 27 sequences was used to test for recombination in the 1948-1955 alignment using the Genetic Algorithm for Recombination Detection. A total of 696 breakpoint placement models were evaluated using the corrected Akaike Information Criterion (AICc) to compare model fit, where a smaller AICc corresponds to a better supported model.

Conserved-domain search and conservation profiling

The same sequence set was submitted to Batch CD Search (NCBI Conserved Domain Database). CD Search works on protein queries and most of the submissions in the nucleotide category 24 out of 31 were processed and none returned an annotated conserved domain in the batch summary. A WebLogo nucleotide sequence logo was also created over the alignment to display the conservation at the position without using the CD Search step.

3 Results

Phylogenetic relationships

The ten Pakistani CTV-CP isolates were not grouped together in a single exclusive cluster in the PhyML tree (Fig. 2). The Pakistani isolates clustered tightly together with very short branch lengths (very recent shared ancestry or limited spread of a dominant strain of this virus) while others Nowshera43, Sawabi10, Sargodha86, Haripur41, Islamabad41 were more distant and interspersed with GenBank CP (SY550 ,SY553) and in some cases with host EST branches alignment noise due to the very different origin of these ESTs as compared to the CTV origin rather than true viral relatedness. The CTV ORF1a genotype markers (B3, D1 ,D4) the three unrelated citrus viruses (CLBV, CiBSV, CiLV N) formed clearly separated lineages of longer branches as expected due to the different genome region (ORF1a) or the different virus species.

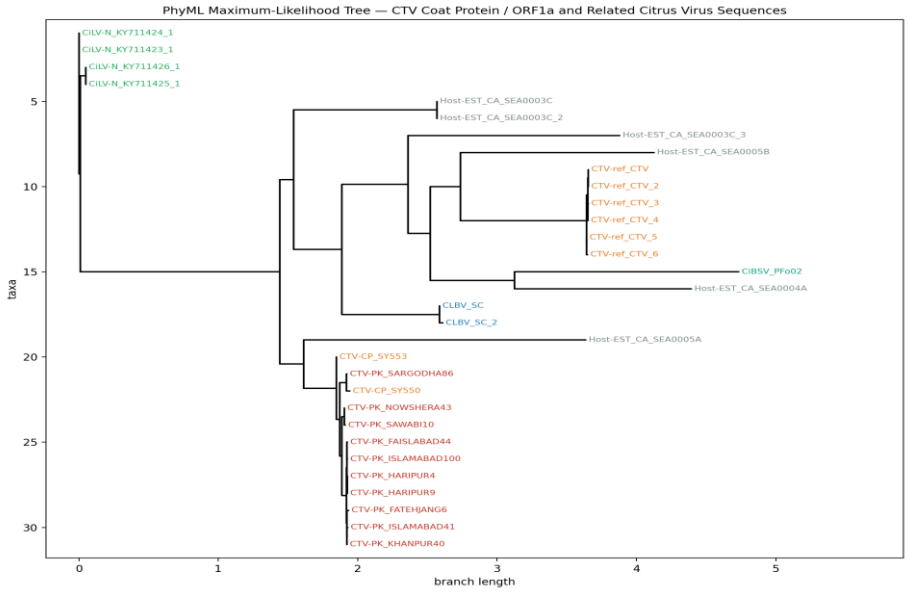

Figure 2. PhyML maximum-likelihood tree of the 456-bp BMGE-trimmed alignment. Tip labels are shortened and colour-grouped, red = Pakistani CTV CP isolates (this study), orange = other CTV CP references, purple = CTV ORF1a genotype markers, grey = host (Citrus) EST cDNA clones, blue teal/green = CLBV, CiBSV and CiLV N respectively.

Pairwise nucleotide identity among Pakistani isolates

The nucleotide identity of the ten Pakistani CTV CP isolates from the BMGE trimmed alignment was 90.8% to 99.8% with most of the pairs being above 96% identity (Fig. 3). This range suggests that the Pakistani isolates are definitely a population of the CTV but are not clonal at least one pair of isolates is on average about 9% different when aligned on the aligned sites which is a characteristic of the genetic difference between different CTV genotypes rather than simple geographic drift of one strain.

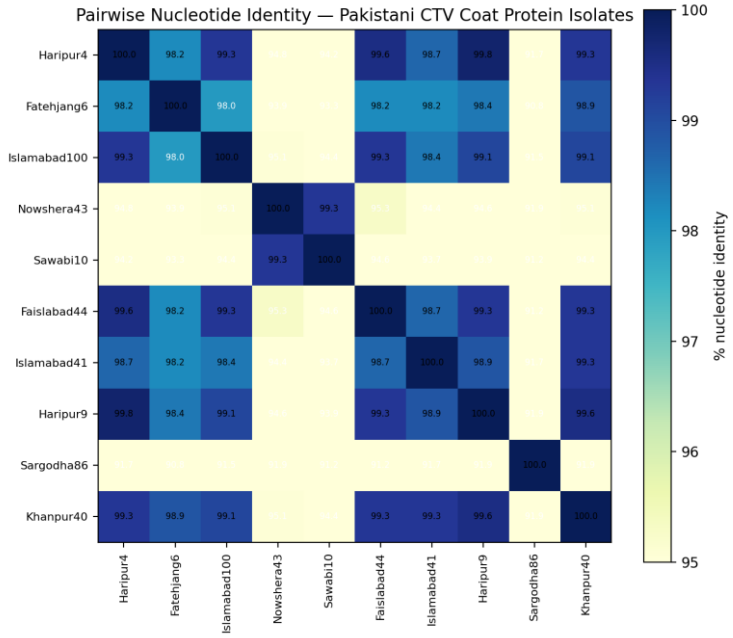

Figure 3. Pairwise nucleotide identity (%) among the ten Pakistani CTV coat protein isolates computed on the BMGE trimmed 456 bp alignment.

Recombination screening

GARD strongly supported a model that includes a single recombination break point at nucleotide position 259 of the 456-bp alignment compared to a model that assumes a single tree for the entire region (ΔAICc = 1197.25 Table 2). This is a very large AICc improvement (above the convention ΔAICc > ~2–10) and suggests that both 5′ and 3′ parts of the alignment analysed (relative to the position 259) are informative of a different topology branch length which is the classical signature of a historical recombination event template switching during replication or strong region-specific selective pressure acting differently on either side of the point 259.

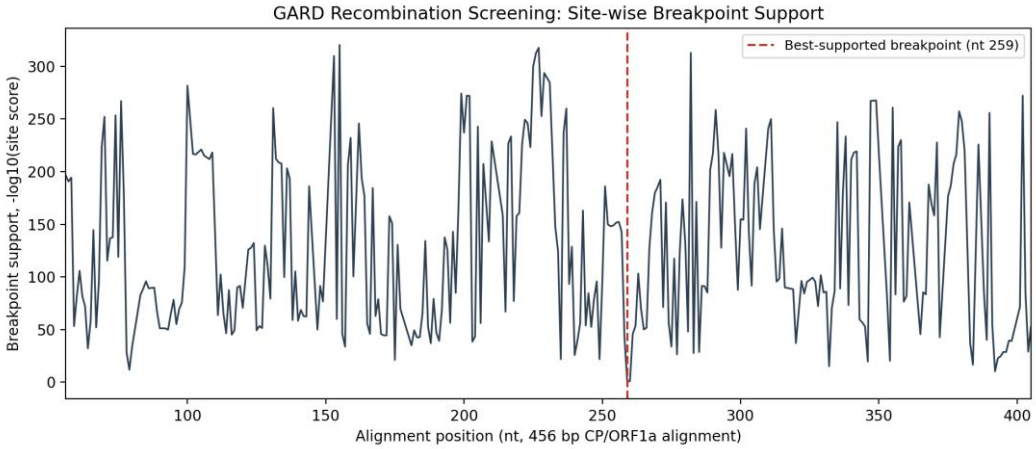

Figure 4. Site wise breakpoint support across the alignment from the GARD screen with the single best supported breakpoint marked at nucleotide position 259 (456 bp alignment 27 sequences 696 candidate models compared).

Table 2. Model comparison from the GARD recombination screen.

Metric	No-breakpoint (single-tree) model	Best GARD model (1 breakpoint, nt 259)
AICc	12082.69	10885.43
ΔAICc vs. single tree	–	1197.25 (strongly favoured)
Candidate models screened	696 total model topologies evaluated across the 456-bp alignment	

Sequence Conservation and Conserved-domains search.

The WebLogo nucleotide conservation profile (Fig. 5) depicts a mosaic of conserved and variable regions with many of the tall single letter stacks representing regions with synonymous constraint in the coding positions which are generally well conserved and variable columns generally located outside the regions flanking the GARD identified breakpoint. As expected Batch CD Search on the same sequence set was not informative at the nucleotide level 24 31 CD Search submissions were ignored as nucleotide query and shorter queries that run did not yield a domain that could be annotated to the Superfamily level in the results summary. This is not a biological discovery but a tool input limitation (CD Search requires protein or well formed ORF translated queries) and is considered here as limitation rather than a discovery.

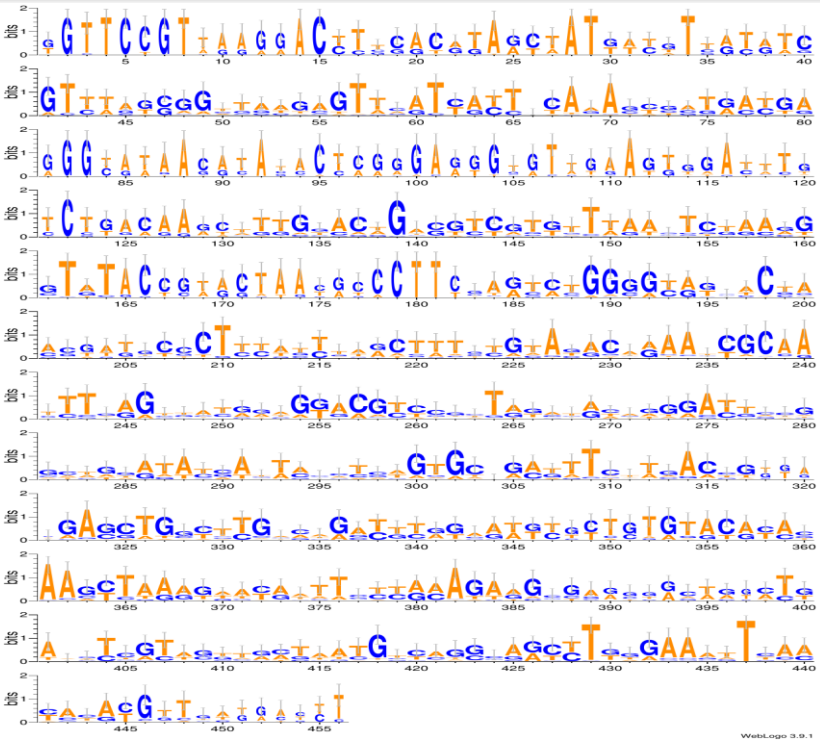

Figure 5. WebLogo nucleotide conservation profile across the 456 bp alignment (bit height indicates positional conservation A/T/G/C stacked by relative frequency).

Table 3. GenBank accessions used in this analysis.

Isolate	Country	Accession	Length (bp)
Haripur4	Pakistan	MF850336.1	672
Fatehjang6	Pakistan	MF850335.1	672
Islamabad100	Pakistan	MF850334.1	672
Nowshera43	Pakistan	MF850333.1	672
Bhan2	India	GQ475572.1	672
Akt1	India	GQ475571.1	672
P14	India	GQ475570.1	672
P13	India	GQ475569.1	672
CTV-0002	China	AJ518841.1	672
CHN-N25	China	EF517336.1	672
CHN-N4	China	EF517335.1	672
CHN-N3	China	EF517334.1	672
L12175	USA	L12175.1	672

The identity of the Pakistani isolates compared to each other at the pairwise level is shown .

4 Discussion

This comparative study shows significant genetic diversity between the isolates of Citrus tristeza virus (CTV) from Pakistan and suggests that the sequences examined do not support a simple evolutionary history by a single clone. A maximum likelihood phylogenetic tree was reconstructed using ten Pakistani coat protein (CP) isolates after alignment using MAFFT and BMGE in the curated 456-bp alignment which showed that the Pakistani CP isolates were not monophyletic. The identities of the two sequences were 90.8% to 99.8% which showed high level of sequence variation amongst the isolates from varied citrus growing areas in Pakistan.

A lack of a single Pakistani specific clade indicates that the sampled isolates might be from multiple evolutionary lineages and not a genetically uniform population. A few Pakistani

isolates were observed on separate branches or close to the CTV reference sequences from other geographical regions in the maximum likelihood tree. It is speculated that the historical introduction local diversification or the diffusion of genetically distinct CTV lineages in the Pakistani citrus production systems may have led to such a phylogeography pattern. However the interpretation made should be considered as provisional because this data is opportunistic from the GenBank data and is not a geographically balanced population sampling scheme. For determining the geographical structure and epidemiological distribution of individual CTV genotypes a more comprehensive sampling of the entire citrus growing areas of Pakistan would be necessary.

The significance of the present phylogeny is related to the fact that the comparative sequence panel is heterogeneous. The dataset included the ten Pakistani CTV CP isolates as well as other CTV CP sequences other virus sequences that infect citrus (CTV ORF1a genotype marker fragments) and EST cDNA sequences from citrus host.

The resulting tree is most informative for showing a broad separation of sequences and is used to provide computational context for the Pakistani CTV sequences. It is not intended to be a complete picture of the global phylogeography of CTV. A comparative analysis of a balanced number of geographically matched isolates from Pakistan, India, China, Brazil, USA and other major citrus growing regions would produce more convincing evidence of geographical structuring and genotype distribution. The evidence from these thirteen CP sequences is also congruent with and supported by the latest field-based evidence in Pakistan. Previous serological survey (Khan et al. 2022) and independent ELISA monitoring survey (Shafique et al. 2024a) carried out in Sargodha district revealed that CTV incidence as well as severity varied significantly among cultivars and among tehsils Mosambi being the most affected and Kot Momin being the most severely affected locality. This coupled with the current molecular re analysis indicates that the genetic diversity of both the CTV variants in circulation and the epidemiological pressure they put on is still mounting within the major citrus belt of Pakistan as well as in a recent review of the emergence and management of CTV in Pakistan (Shafique et al. 2024b).

5 Conclusion

Significant genetic diversity was observed between the ten different CTV coat protein isolates from Pakistan with 90.8% to 99.8% nucleotide identities between them. The maximum likelihood phylogeny based on the alignment built using MAFFT BMGE did not support all the Pakistani isolates belonging to a single exclusive clade which indicated the presence of genetically diverse CTV lineages in the sampled citrus material in Pakistan. In addition the GARD analysis revealed a candidate recombination breakpoint close to position 259 that suggests that multiple parts of the examined sequence might have evolved separately.

These results are generally consistent with the genetic complexity, lineage diversity and recombination potential of CTV populations that have been established (Harper, 2013, Harper et al., 2010, Roy et al., 2013, Roy et al., 2020). The recombination signal detected should be considered a candidate signal for the rebound and not a biological signal for a recombinant isolate. Similarly the nonexistence of a single Pakistani clade is not a proof of geographical population structure as the GenBank derived data was not a geography matched sampling set.

In general, the present study offers a molecular reference of CTV sequence variability in Pakistan and illustrates the usefulness of performing MAFFT BMGE alignment PhyML maximum likelihood phylogenetics and GARD recombination screening in the exploratory sequence analysis of viruses. More representative number of geographically dispersed Pakistani isolates with completed CP sequences and ideally whole genome sequence should be included in future research. The integration of these data with various independent recombination detection methods and epidemiological data would help to characterize the CTV genotype better understand evolutionary relationships,

recombination patterns and geographical dissemination in Pakistan.

6 Reference

Bashir, S., Mubeen, M., Iftikhar, Y., Zeshan, M. A., Safdar, H., Shabbir, M. A., ... & Sajid, A. (2026). Monitoring of the citrus greening disease (Huanglongbing) in relation to environmental factors in Pakistan. *Plant Protection News (Vestnik Zashchity Rasteniy)*, (1), 98-103.

Bashir, S., Fatima, R., Shabbir, M. A., Fareed, M. A., Bakhtawar, F., Basheer, S., ... & Aslam, H. (2025). Aphis gossypii transmitted Citrus Tristeza Virus: A Major Threat to Production of Citrus with Emphasis. *Medical and Life Sciences*, *4*(2), 48-54.

Shafique, T., Shabbir, M. A., Rehman, A. U., Saeed, M. H., Haider, S. A., Fareed, M. A., & Bilqees, I. (2024). Citrus Tristeza Virus in Pakistan: Unraveling its Emergence, Evolution, and Impact on the Citrus Industry. *Medical and Life Sciences*, *3*(1), 1-8.

Shafique, T., Shabbir, M. A., Rehman, A. U., Saeed, M. H., Haider, S. A., Fareed, M. A., & Bilqees, I. (2024). Citrus Tristeza Virus in Pakistan: Unraveling its Emergence, Evolution, and Impact on the Citrus Industry. *Medical and Life Sciences*, *3*(1), 1-8.

Naveed, K., Shafiq, M. R., & Raza, A. (2023). Citrus tristeza virus: its research status and future prospects in Pakistan.

Wang, C., Chen, C., Chen, Y., Zhong, K., & Yi, L. (2023). Bayesian phylodynamic analysis reveals the evolutionary history and the dispersal patterns of citrus tristeza virus in China based on the p25 gene. *Virology Journal*, *20*(1), 223.

Bester, R., Massart, S., Fontdevila Pareta, N., Khalili, M., Maachi, A., Rivarez, M. P. S., ... & Petter, R. (2023). Managing the deluge of newly discovered plant viruses and viroids: an optimized scientific and regulatory framework for their characterization and risk analysis.

Tyagi, S., Kabade, P. G., Gnanapragasam, N., Singh, U. M., Gurjar, A. K. S., Rai, A., ... & Singh, V. K. (2023). Codon usage provide insights into the adaptation of rice genes under stress condition. *International Journal of Molecular Sciences*, *24*(2), 1098.

Karasev, A.V. et al. (1995) Complete sequence of the citrus tristeza virus RNA genome. Virology 208, 511–520.

Rubio, L. et al. (2001) Genetic variation of Citrus tristeza virus isolates from California and Spain. Phytopathology 91, 762–768.

Kosakovsky Pond, S.L. et al. (2006) GARD: a genetic algorithm for recombination detection. Bioinformatics 22, 3096–3098.

Guindon, S. & Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52, 696–704.

Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7. Molecular Biology and Evolution 30, 772–780.

Criscuolo, A. & Gribaldo, S. (2010) BMGE: a software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evolutionary Biology 10, 210.

Marchler-Bauer, A. et al. (2017) CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Research 45, D200–D203.

Crooks, G.E. et al. (2004) WebLogo: a sequence logo generator. Genome Research 14, 1188–1190.

Crooks, G. E., Hon, G., Chandonia, J.-M., & Brenner, S. E. (2004). WebLogo: A sequence logo generator. *Genome Research*, *14*(6), 1188–1190. https://doi.org/10.1101/gr.849004

Dawson, W. O., Garnsey, S. M., Tatineni, S., Folimonova, S. Y., Harper, S. J., & Gowda, S. (2013). *Citrus tristeza virus*–host interactions. *Frontiers in Microbiology, 4*, 88. https://doi.org/10.3389/fmicb.2013.00088

Ghori, N.-U.-H., Hayat, M. Q., & Davino, S. W. (2016). Genetic diversity and evolutionary analysis of *Citrus tristeza virus* p20 gene in Pakistan: Insights into the spread and epidemiology. *Advancements in Life Sciences, 3*(3), 75–82.