Malaysian Journal of Microbiology, Vol 18(1) 2022, pp. 79-92
DOI: http://dx.doi.org/10.21161/mjm.211276

Malaysian Journal of Microbiology
Published by Malaysian Society for Microbiology
(In SCOPUS since 2011)

Cucumber mosaic virus: Global genome comparison and beyond

Leonard Whye Kit Lim*, Ing Mee Hung and Hung Hui Chung*

Faculty of Resource Science and Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia.
Email: lwkl1993@gmail.com; hhchung@unimas.my

Received 17 August 2021; Received in revised form 20 September 2021; Accepted 20 December 2021

ABSTRACT

Aims: The cucumber mosaic virus (CMV) is categorized under the genus *Cucumovirus* and family Bromoviridae. This virus is known to infect over 1200 plant species from 100 families, including ornamental and horticultural plants. In this study, we pioneered a global genome comparison to decipher the unknown orchestrators behind the virulence and pathogenicity of CMV via the discovery of important single nucleotide polymorphic markers.
Methodology and results: As a result, the genome size was found to be a potential preliminary country-specific marker for South Korea and the GC content can be utilized to preliminarily differentiate Turkey isolates from the others. The motif analysis as well as whole genome and coat protein phylogenetic trees were unable to form country-specific clusters. However, the coat protein haplotype analysis had successfully unconcealed country-specific single nucleotide polymorphic markers for Iran, Turkey and Japan isolates. Moreover, coat protein modelling and gene ontology prediction depicted high conservation across CMV isolates from different countries.
Conclusion, significance and impact of study: The country-specific single nucleotide polymorphic markers unearthed in this study may provide significant data towards the profiling of varying virulence and pathogenicity of CMV across the globe in time to combat the yield loss driven by this virus thru the most efficacious biological control measures in the future.

Keywords: Cucumber mosaic virus, genome comparison, country-specific SNPs, haplotype, coat protein

INTRODUCTION

The cucumber mosaic virus (CMV) is a *Cucumovirus* genus member grouped under the family Bromoviridae. It has the capability to infect over 1200 plant species from plant families exceeding 100. The host range of this virus is not limited to ornamental plants only but it also infects horticultural plants (Zitter and Murphy, 2009) and most virus host plants are perennial or biennial in nature (Draegar, 2016).

The 29 nm diameter CMV virions are icosahedral in nature and they possess 180 subunits of a single capsid protein (CP) as well as 18% RNA (Jacquemond, 2012). This virus composed of a sum of three single-stranded positive-sense RNAs, with names 1 to 3, in the order of diminishing length. The monocistronic RNA1 encodes for protein 1a, in which the N-terminal encompasses the putative methyltransferase domain whereas the C-terminal houses the helicase motif (Jacquemond, 2012). The RNA2 encodes for the larger 2a protein and smaller 2b protein whereby the 2a protein contains GDD motif pivotal to RNA-dependant RNA polymerase (RdRp) functioning (Jacquemond, 2012). The 2b protein disrupts the RNA interfering (RNAi) pathway of the host. The RNA3 is bicistronic and it encodes for the coat protein (CP) as well as the movement protein (MP) or also known as the 3a protein (Jacquemond, 2012). The total lengths of RNAs differ across different strain groupings and strains and this has resulted in varying genome sizes.

The CMV genome sizes range from 8500 to 8700 bp and it has been sequenced from various host species like radish, turnip weed, wild cabbage, tobacco, musk melon, field pumpkin, tomato, cucumber, spinach, Indian shot, black mustard seed, balsam, Adzuki bean, fungus, naranjilla and Jew's melon, as well as countries like Iran, Japan, Turkey, South Korea, Brazil, China, Colombia, France, Germany, Poland and USA (GenBank, 2021). A sum of 249 full genomes were found deposited in the public GenBank database to date but we only select the 40 genomes published in renowned high impact journal articles to ensure high fidelity and confidence on the country of origin and host species profiles reported (GenBank, 2021). The commonest host species is the radish, followed by the field pumpkin. The three major CMV genome sequence contributors are researchers from Iran, Japan and Turkey (GenBank, 2021). With the

*Corresponding author

ISSN (print): 1823-8262, ISSN (online): 2231-7538

Malays. J. Microbiol. Vol 18(1) 2022, pp. 79-92
DOI: http://dx.doi.org/10.21161/mjm.211276

availability of the genomes thus far, yet there is a lack of global full genome comparison to date, to the best of our knowledge.

The whole genome comparison would aid greatly in assisting us to view the full picture of the genomic landscape variation across countries and host species. Added to our advantage is that the end of the cost and labor-intensive genome sequencing era as the next generation sequencing has created wonders in the exponential growth of the genomes being sequenced at an unprecedented pace. Whole genome contrasting across biogeographical localities has been proven to provide significantly high resolution right to the single nucleotide polymorphism (SNP) level via genomics-based haplotype analyses (Lim *et al.*, 2021) and the identified SNPs are vital contributors towards varying host preferences and disease symptoms (Katsir *et al.*, 2018). Hence, we aimed to conduct genome-wide comparisons across different host species and countries to decipher the undiscovered variations in virulence besides dissecting on the coat protein of CMV. These data could significantly assist future studies involving the CMV virulence pathway comprehension and characterization.

MATERIALS AND METHODS

Characterization of genomes

A sum of 40 complete CMV genome sequences (we select only the ones with data published in renowned journal articles to ensure high fidelity on the country of origin and host species profiles reported) were retrieved from the public GenBank database. All CMV full genomes were deposited into the GenBank database as three separate genome fragments with three assigned accession numbers. Therefore, these genome fragments were concatenated prior to the full genome analyses. A viral genome map was constructed using the full CMV genome of Iran IRN-BRE5 isolate from host *Raphanus sativus* (radish) (GenBank accession numbers: LC066459.1, LC066460.1, LC066461.1) via CGView (Grant *et al.*, 2012). The GC contents and genome sizes were enumerated using MEGA X (Kumar *et al.*, 2018). The mVISTA software (Frazer *et al.*, 2004) was utilized to conduct the comparative and divergence analyses across selected CMV genomes isolated from different host species and country, in Shuffle-LAGAN mode. Maximum likelihood phylogenetic tree was constructed for both full genomes and coat protein sequences employing MEGA X (Kumar *et al.*, 2018) with the Tamura-Nei model and 1000 bootstrap replication. Haplotype analysis was conducted on coat protein sequences to identify country-specific SNP markers.

Characterization and modelling of coat protein

The coat protein was chosen for the protein characterization and modelling because it has been discovered to play indispensable role in stunt disease phenotypic expression (Qiu *et al.*, 2018). The coat protein profiles (amino acid length, molecular weight and isoelectric point (pI)) were compared across Japan, Turkey, Iran and South Korea isolates using ProtParam (Gasteiger *et al.*, 2005). These proteins were further characterized via ScanProsite (de Castro *et al.*, 2006), NCBI conserved domain search (Yang *et al.*, 2020) and MEME suite (Bailey *et al.*, 2015). CELLO server version 2.5 was utilized for subcellular localization prediction whereas Dali (Holm *et al.*, 2008) was employed to unearth top five structural neighbors of the protein. Phyre2 server (Kelley *et al.*, 2015) was used to model the protein in three dimension. The Ramachandran plot was used to verify the integrity of protein structure predicted. The binding sites of the protein were predicted using 3DLigandSite (Wass *et al.*, 2021). The examination of the amino acid variation sites was also performed.

RESULTS AND DISCUSSION

Genome overview

To date, out of the selected 40 CMV isolates with full genome sequences, the most abundant host species is the radish (*R. sativus*) with composition of 37.5%. These CMV genomes isolate from radish are sequenced by researchers from Iran and Japan only. In fact, these two countries are two major CMV genome reservoirs accounting 35% (Japan) and 25% (Iran) of the total CMV genomes sequenced. Other host species involved are field pumpkin, spinach, tobacco, tomato, wild cabbage, turnip weed, muskmelon, Black mustard seed, Adzuki bean, common bean, Indian shot, balsam, Jew's melon and cucumber, isolated from countries like Poland, Turkey, Brazil, South Korea, France, Germany, China, Colombia and USA (Table 1).

The viral genome map of Iran IRN-BRE5 isolate (GenBank accession numbers: LC066459.1, LC066460.1, LC066461.1) (Figure 1) depicted five major coding regions. A sum of two non-coding regions were detected within the genome, one is located before the movement protein and one is found before the coat protein sequence. Replicase gene was observed at around 200 to 3000 bp. The RNA dependant RNA polymerase gene was revealed at 3400 to 5900 bp, which it overlaps with the viral suppressor gene for RNA silencing at 5700 to 6000 bp. This viral genome map provides the general genome structure of the CMV genome which is conserved across all isolates. Nevertheless, the GC contents and genome sizes are different across different country and host species.

Genome size and GC content are potential preliminary country-specific markers for South Korea and Turkey respectively

The genome sizes and GC contents were compared across all 40 CMV genomes as shown in Figure 2. The genome lengths ranged from 8500 to 8651 bp whereas the GC contents are all 46.3 ± 0.3%. Looking at the whole genome comparison (Figure 2A), the genome sizes and

 ISSN (print): 1823-8262, ISSN (online): 2231-7538

Malays. J. Microbiol. Vol 18(1) 2022, pp. 79-92
DOI: http://dx.doi.org/10.21161/mjm.211276

Table 1: The general information on all CMV genome sequences used in this study.

GenBank accession numbers			CMV isolate name	Host organism	Country of origin
RNA1	RNA2	RNA3			
LC066459.1	LC066460.1	LC066461.1	IRN-BRE5	*Raphanus sativus* (Radish)	Iran
LC066483.1	LC066484.1	LC066485.1	IRN-WRN8		
LC066486.1	LC066487.1	LC066488.1	IRN-WRN9		
LC066489.1	LC066490.1	LC066491.1	IRN-WRN15		
LC066492.1	LC066493.1	LC066494.1	IRN-WRSh41		
LC066495.1	LC066496.1	LC066497.1	IRN-WRWA5		
LC066408.1	LC066409.1	LC066410.1	IWD041J		Japan
LC066411.1	LC066412.1	LC066413.1	KD302J		
LC066414.1	LC066415.1	LC066416.1	MD965J		
LC066429.1	LC066430.1	LC066431.1	NND454J		
LC066435.1	LC066436.1	LC066437.1	OD291J		
LC066438.1	LC066439.1	LC066440.1	OGI11X2		
LC066441.1	LC066442.1	LC066442.1	OGI12J3		
LC066450.1	LC066451.1	LC066452.1	TYD083J		
LC066405.1	LC066406.1	LC066407.1	GFD560J		
LC066507.1	LC066508.1	LC066509.1	TUR83	*Rapistrum rugosum* (Turnip weed)	Turkey
LC066510.1	LC066511.1	LC066512.1	TUR84		
LC066513.1	LC066514.1	LC066515.1	TUR86		
LC066516.1	LC066517.1	LC066518.1	TUR89		
MG882753.1	MG882754.1	MG882755.1	CMV 21	*Cucurbita pepo* (Field pumpkin)	Poland
MH782238.1	MH782239.1	MH782240.1	SqSh		Iran
GU327366.1	GU327367.1	GU327368.1	Z1		South Korea
KY886409.1	KY886410.1	KY886411.1	CMV-SP	*Spinach oleracea* (Spinach)	Brazil
LC066420.1	LC066421.1	LC066422.1	MS655J		Japan
AB188234.1	AB188235.1	AB188236.1	CM95	*Nicotiana tabacum* (Tabacco plant)	Japan
LC363917.1	LC363918.1	LC363919.1	Co-46	*Corchorus olitorius* (Jew's mallow)	
AB188231.1	AB188232.1	AB188233.1	Fuka4-4	*Cucumis sativus* (Cucumber)	
LC066432.1	LC066433.1	LC066434.1	NRB648J	*Brassica juncea* (Black mustard seed)	
LC066477.1	LC066478.1	LC066479.1	IRN-Tlm1	*Impatiens balsamina* (Balsam)	Iran
LC066480.1	LC066481.1	LC066482.1	IRN-TVRa26	*Rapistrum rugosum* (Turnip weed)	
MH782235.1	MH782236.1	MH782237.1	MeEs	*Cucumis melo* (Muskmelon)	
LC381764.1	LC381763.1	LC381757.1	Can	*Canna generalis* (Indian shot)	South Korea
AJ276479.1	AJ276480.1	AJ276481.1	Mf	-	
JX014246.1	JX014247.1	JX014248.1	Va	*Vigna angularis* (Adzuki bean)	
HF572914.1	HF572915.1	HF572916.1	Bn57	*Phaseolus vulgaris* (Common bean)	USA
HE793683.1	HE793684.1	Y18137.1	I17F	*Solanum lycopersicum* (Tomato)	France
KX525731.1	KX525735.1	KX525739.1	PV-0036	-	Germany
MG025947.1	MG025948.1	MG025949.1	Rs	*Rhizotocnia solanum* (Fungus)	China
MG696854.1	MG696855.1	MG696856.1	San-Vicente 1	*Solanum quitoense* (Naranjilla)	Colombia
LC066498.1	LC066499.1	LC066500.1	TUR4	*Brassica oleracea* (Wild cabbage)	Turkey

GC contents of all Iran and Japan isolates do not differ significantly. The largest genome belongs to that of Japan isolated from Jew's melon (C0-46 isolate) (8651 bp) while the smallest genome is that of Iran MeEs isolated from turnip weed (8501 bp). Upon ANOVA testing, the GC contents of Turkey isolates are significantly higher than those from Iran, Japan and South Korea. Similarly, the genome sizes of South Korea isolates stand out significantly among the four countries. Therefore, the genome size can be a potential preliminary country-specific marker for South Korea and the GC content can be used to preliminarily distinguish Turkey isolates from the others.

Analogous phenomena were observed in the respective genome fragment analyses in this study. For instance, the Turkey isolates GC content of RNA1 genome fragments are statistically significant in contrast to others, determined via ANOVA and post hoc analyses (p<0.05) (Figure 2B). Interestingly, the genome fragment lengths of South Korea isolates stand out statistically (ANOVA and post hoc analyses (p<0.05)) among the four countries (Japan, Iran, Turkey and South Korea) with regards to RNA2 (Figure 2C) and RNA3 (Figure 2D) genome fragments. These findings are as important as the findings from the whole genome because it may provide a cost effective screening in times where the full

 ISSN (print): 1823-8262, ISSN (online): 2231-7538

Malays. J. Microbiol. Vol 18(1) 2022, pp. 79-92
DOI: http://dx.doi.org/10.21161/mjm.211276

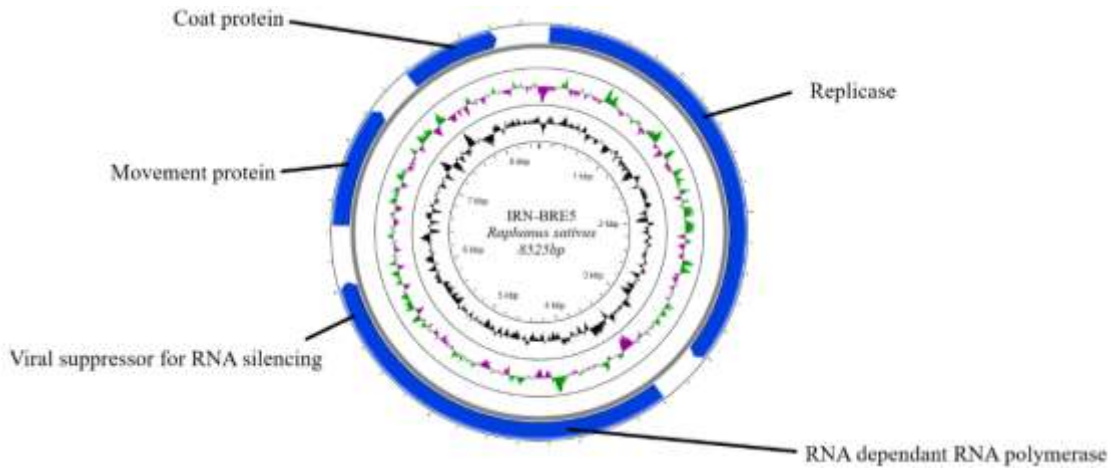


Figure 1: The viral genome map of Iran IRN-BRE5 isolate (GenBank accession numbers: LC066459.1; LC066460.1; LC066461.1). The coding regions were represented by the blue regions, positive GC skew was represented by the green spike, negative GC skew was represented by the purple spike, GC content was represented by the back spike.

genome of an unknown isolate is incomplete and only one of the genome fragment is available. In short, the RNA2 and RNA3 genome fragment lengths is adequate to tell South Korea isolates apart from other countries whereas the GC content of RNA1 genome fragment is enough to distinguish Turkey isolates from the others in times without the need for the full genome sequences.

Motif analysis as well as whole genome and coat protein phylogenetic trees failed to resolve country-specific clusters

An overview of the genome alignment across all four isolates from four different countries were depicted in Figure 3A. At a glance, only the South Korea Z1 isolate is notably different from the other isolates (Figure 3A). A total of five motifs detected across the coat protein genes of the four isolates originating from four different countries (Figure 3B). The motif length, motif distribution pattern, number of motifs and motif position are all conserved across all four isolates from Iran, Japan, Turkey and South Korea correspondingly. The first motif "MDKSESTSAGRNRRRRPRRGSRSAPSSAD" is arginine (R) abundant and it is the highlight within the coat protein gene. This motif is the only functional motifs out of the other motifs found within the CMV coat protein gene (Ng *et al.*, 2000), deemed to play role in binding the virus to plant cells (Zhang *et al.*, 2017). Despite the motif analysis failed to resolve any country-specific clusters, the presence of this motif in all isolates strongly verified that all isolates have high degree of virulence based on the presence of this arginine-rich motif contributing towards the chlorosis induction in old leaves and extreme stunting in host organisms (Zhang *et al.*, 2017).

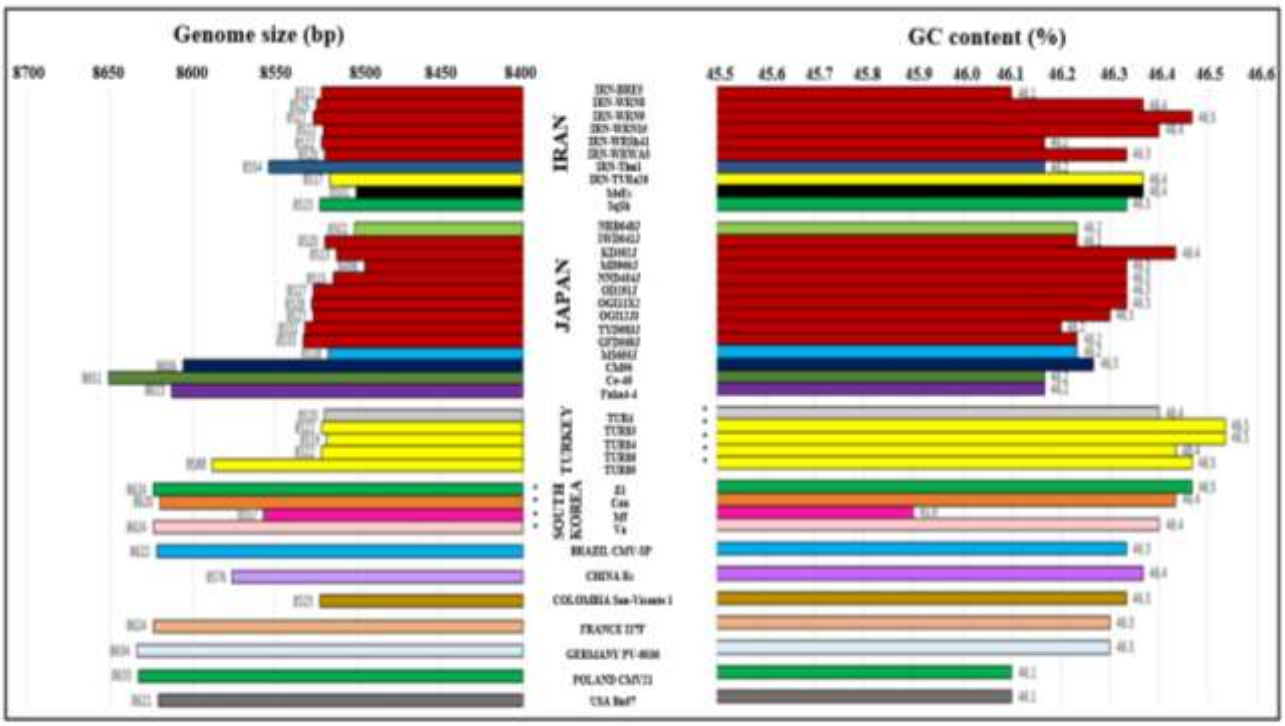
The coat protein and whole genome maximum likelihood phylogenetic trees (Figure 4) showed only two clusters. The same phenomena was observed across individual trees constructed using the respective genome fragments, with regards to the whole genome tree. The first cluster housed the two outgroups, namely tomato mosaic virus OM isolate (GenBank accession number: AB355139.1) and barley yellow mosaic virus BAYMV-Daegu isolate (GenBank accession numbers: MG891924.1, MG891929.1). The second cluster was resided by all 40 isolates used in this study. In short, there is no country-specific clustering formed across all 40 CMV isolates. Similar outcomes were obtained from Kimaru *et al.* (2020) where the isolates they sequenced are clustered with that from the USA (Nouri *et al.*, 2014) and South Korea (Kim *et al.*, 2014). However, Kimaru *et al.* (2020) had not included as many genome sequences as in this study for comparison. Nematollahi *et al.* (2012) did performed simple global comparison phylogenetic analyses but they only utilized a portion of the genome instead of the full genome. It is rare that the whole genome phylogenetic tree is unable to resolve country-specific clusters as this has been proven success in the past (Lin *et al.*, 2013; Lim *et al.*, 2018; Yokono *et al.*, 2018; Lim *et al.*, 2020a; Lim *et al.*, 2020b; Shakya *et al.*, 2020). Hence, there is a need to zoom into the nucleotide level resolution to unearth the dissimilarity across country and host species via haplotype analysis.

Coat protein haplotype analysis revealed country-specific SNP markers for Iran, Turkey and Japan isolates

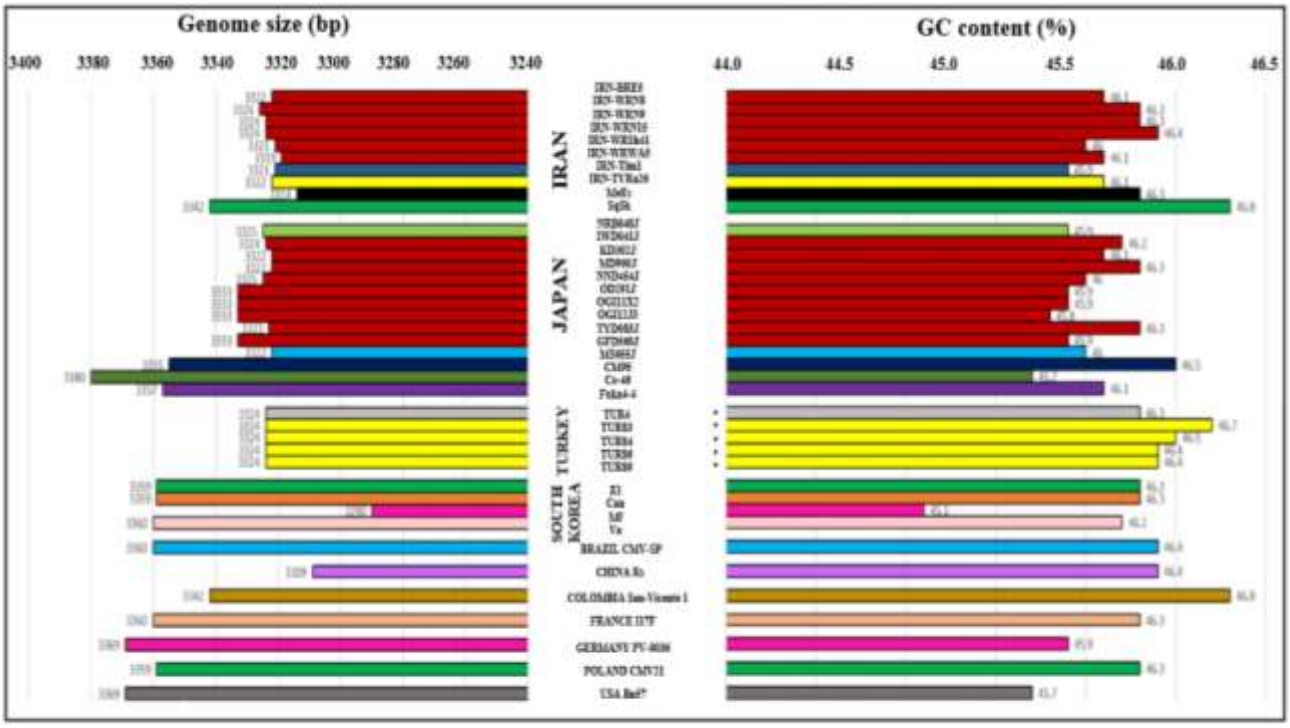
The coat protein haplotype analysis (Tables 2 and 3) revealed that the 33 CMV isolates from Turkey, Japan, South Korea and Iran were grouped into a sum of 27 haplotypes. A total of 51 parsimony SNP sites detected across these isolates and most of them are found within the Japan isolates. Besides, there are several country-specific SNP markers found from the haplotype table. For instance, Site 85 was found to be Japan-specific (with two South Korea exceptions), Site 93 was found to be specific

 ISSN (print): 1823-8262, ISSN (online): 2231-7538

Malays. J. Microbiol. Vol 18(1) 2022, pp. 79-92
DOI: http://dx.doi.org/10.21161/mjm.211276



(A)

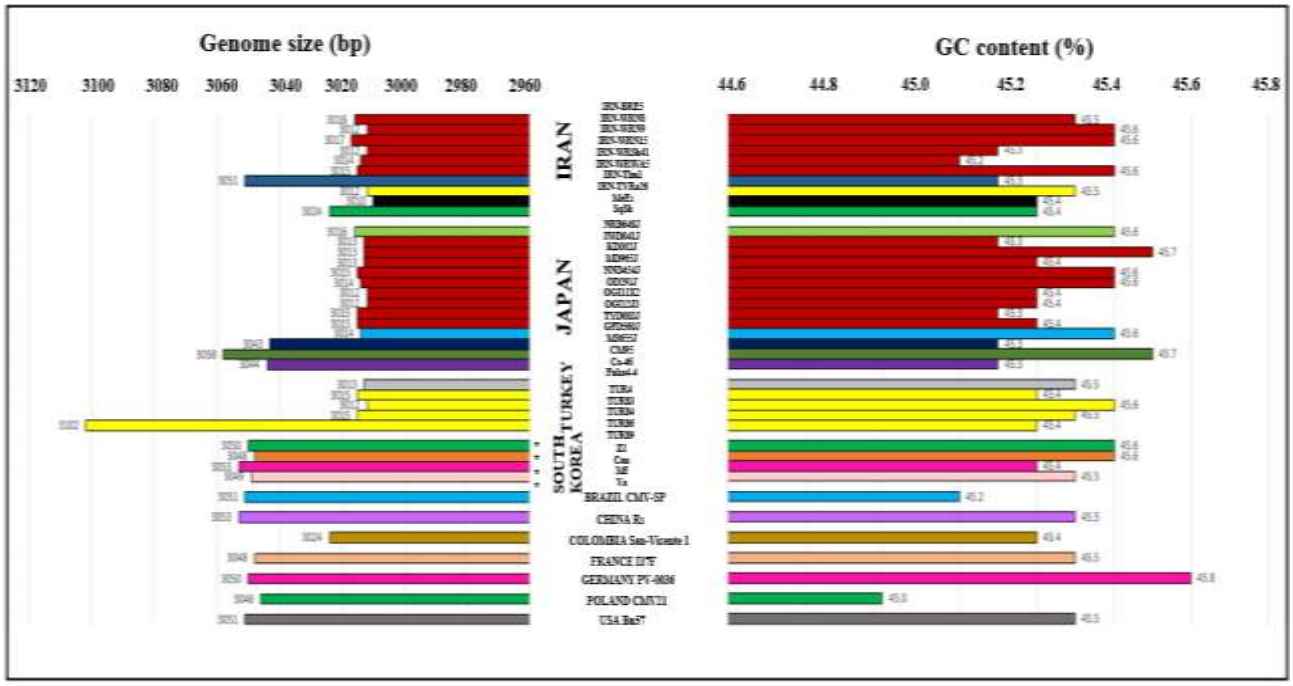


(B)

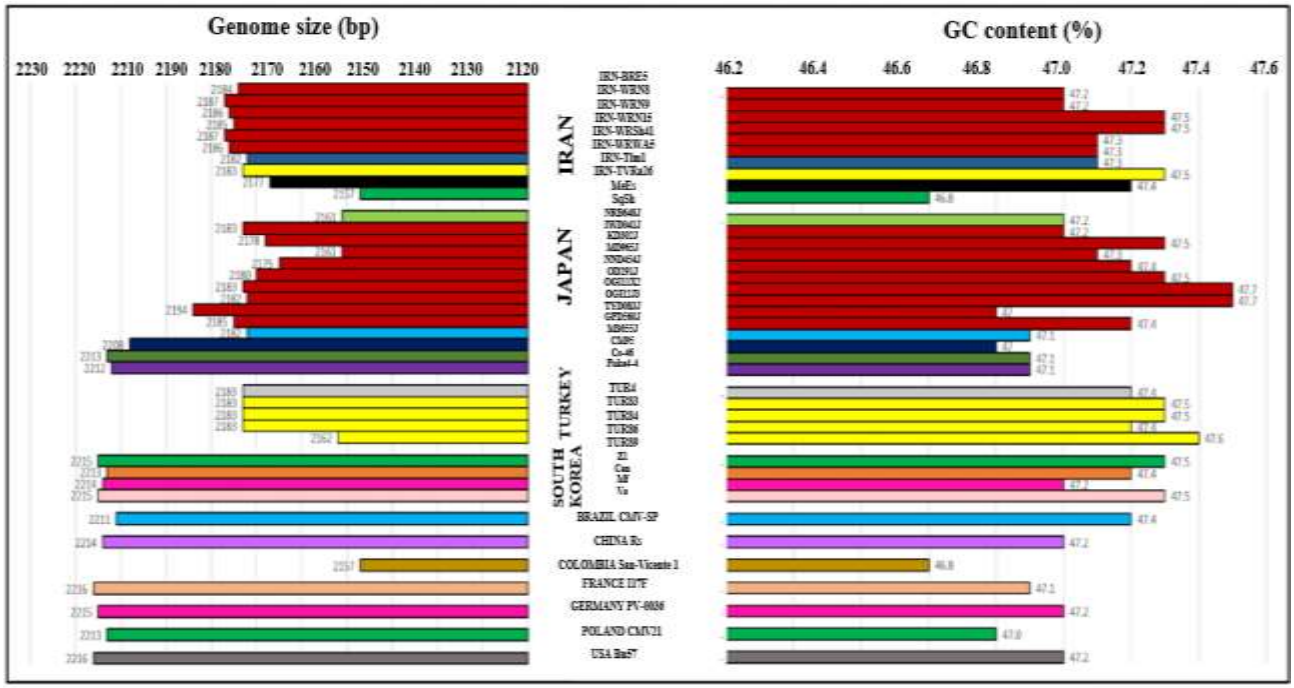
(Continued)

 ISSN (print): 1823-8262, ISSN (online): 2231-7538

Malays. J. Microbiol. Vol 18(1) 2022, pp. 79-92
DOI: http://dx.doi.org/10.21161/mjm.211276



(C)



(D)

Figure 2: The GC contents and genome sizes comparison across (A) full genomes, (B) RNA1 genome fragments, (C) RNA2 genome fragments as well as (D) RNA3 genome fragments. Isolates that are significantly different from others were labelled with a "*", determined via ANOVA and Duncan's post hoc analyses.

 ISSN (print): 1823-8262, ISSN (online): 2231-7538

Malays. J. Microbiol. Vol 18(1) 2022, pp. 79-92
DOI: http://dx.doi.org/10.21161/mjm.211276

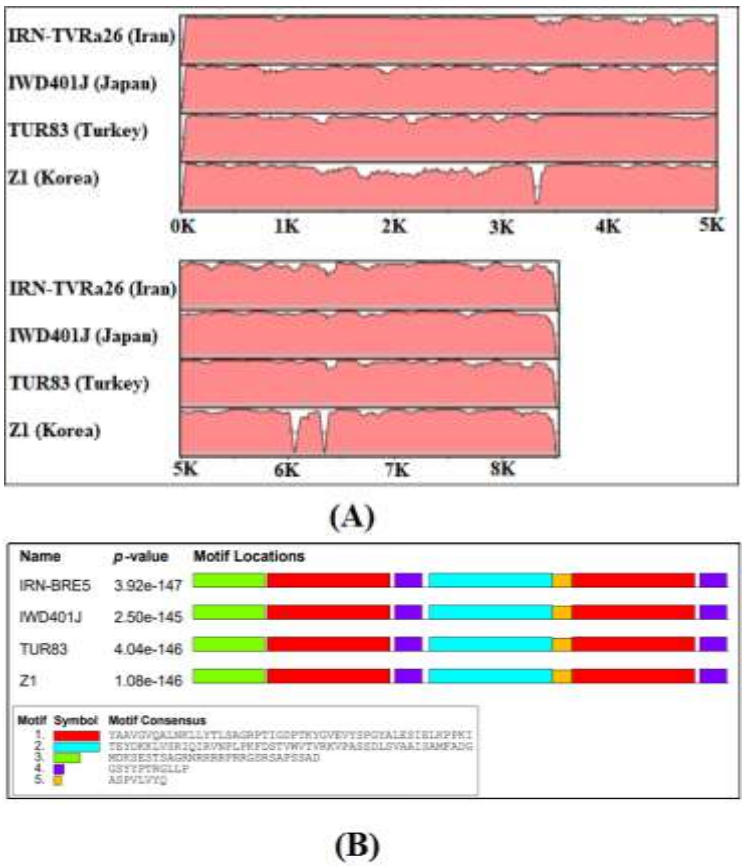


Figure 3: (A) The genome alignment overview visualized across the selected four CMV isolates and (B) the motif analysis of coat proteins of all four selected CMV isolates.

Figure 4: The consensus maximum likelihood phylogenetic tree from trees constructed using coat protein genes and full genomes of CMV isolates and outgroups, with 1000 bootstrap replications.

 ISSN (print): 1823-8262, ISSN (online): 2231-7538

Malays. J. Microbiol. Vol 18(1) 2022, pp. 79-92
DOI: http://dx.doi.org/10.21161/mjm.211276

Table 2: The haplotype analysis across CMV coat proteins for variable sites 12 to 277.

Haplotypes for CMV coat protein	Variable site																										
	12	33	36	57	75	76	84	85	93	114	126	132	141	142	159	186	189	191	192	197	198	204	222	237	249	261	277
H1 IRN-BRE5	T	T	C	T	C	C	C	G	C	G	A	T	G	T	T	T	G	G	T	A	A	A	T	A	T	T	T
H2 IRN-WRN8, IRN-WRN9, IRN-WRSh41	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
H3 IRN-WRN15	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
H4 IRN-WRWA5	.	C	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.
H5 IRN-TVRa26	.	C	T	.	.	.	.	.	T	.	.	.	A	G	.	.	A	.	C	G	.	G	.	.	.	.	.
H6 IRN-Tlm1	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	C
H7 MeEs	.	C	.	.	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	.	.	.	.	.	.
H8 SqSh	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.
H9 IWD041J	.	.	.	.	T	T	.	T	T	.	.	C	A	G	.	.	.	.	.	.	.	.	.	.	C	.	.
H10 KD302J	.	.	.	.	.	T	G	T	T	.	.	.	A	G	.	.	A	.	C	.	.	.	.	.	C	.	.
H11 TYD083J	.	.	.	.	T	T	T	T	T	.	.	.	A	G	.	.	A	.	C	.	.	G	C	.	.	C	.
H12 NND454J	.	.	.	.	.	T	.	T	T	.	.	.	A	G	.	.	A	.	C	.	.	G	.	.	C	.	C
H13 CM95, Fuka4-4	.	.	.	.	.	T	.	T	T	.	.	C	A	G	.	C	A	C	G	.	.	G	.	.	.	.	.
H14 Co-46	.	.	.	.	.	T	.	T	T	A	.	.	A	G	.	.	.	.	.	.	.	.	C	.	C	C	.
H15 MS5655J	.	.	.	.	.	T	.	.	.	.	.	.	.	G	C	.	A	.	C	.	.	G	.	.	C	C	.
H16 NRB648J	.	.	.	.	.	.	.	T	T	.	.	C	A	G	.	.	.	.	.	.	.	.	.	G	C	.	.
H17 MD965J	.	.	T	C	T	T	.	T	T	A	.	.	A	G	.	.	A	.	C	.	.	.	.	.	C	.	.
H18 OD291J	C	.	.	.	.	T	G	T	T	.	.	.	A	G	.	.	A	.	C	.	.	.	.	.	C	C	.
H19 OGI11X2, OGI12J3	C	.	.	.	.	T	.	T	T	.	.	.	A	G	C	.	A	.	C	.	G	.	.	.	C	C	.
H20 GFD560J	C	.	.	.	.	T	.	T	T	.	.	.	A	G	C	.	.	.	.	.	.	.	C	.	C	C	.
H21 TUR4	.	.	.	.	.	.	.	.	T	.	.	.	A	G	.	.	A	.	C	G	.	G	.	.	.	C	C
H22 TUR84	.	.	.	.	.	.	.	.	T	.	G	.	.	G	.	.	A	.	C	G	.	G	.	.	.	C	C
H23 TUR83, TUR86, TUR89	.	.	.	.	.	.	.	.	T	.	G	.	A	G	.	.	A	.	C	G	.	G	.	.	.	C	C
H24 Can	.	.	.	.	.	.	.	T	T	.	.	C	A	G	.	.	A	.	C	.	.	G	.	.	.	.	.
H25 Mf	.	.	.	.	T	.	.	.	T	.	.	.	A	G	.	.	A	.	C	G	.	.	.	.	C	C	C
H26 Va	.	.	.	C	T	.	.	.	T	.	.	.	A	.	C	.	.	.	C	.	.	.	.	.	.	.	.
H27 Z1	.	.	.	.	.	.	.	T	T	.	.	C	A	G	.	.	.	.	C	.	.	.	.	.	C	.	.

Coloured boxes are used to distinguish between isolates from different countries (Blue: Iran; Yellow: Japan; Green: Turkey; Red: South Korea).

ISSN (print): 1823-8262, ISSN (online): 2231-7538

Malays. J. Microbiol. Vol 18(1) 2022, pp. 79-92
DOI: http://dx.doi.org/10.21161/mjm.211276

Table 3: The haplotype analysis across CMV coat proteins for variable sites 288 to 659.

Haplotypes for CMV coat protein	Variable site																							Number of SNP markers
	288	294	309	354	363	381	412	432	450	468	471	483	489	499	537	567	582	598	600	616	639	645	657	659	
H1 IRN-BRE5	A	A	A	T	C	T	G	T	G	T	T	C	T	T	C	C	A	C	A	G	G	G	A	A	0
H2 IRN-WRN8, IRN-WRN9, IRN-WRSh41	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0
H3 IRN-WRN15	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
H4 IRN-WRWA5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2
H5 IRN-TVRa26	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	1
H6 IRN-Tlm1	.	G	G	.	.	.	.	C	.	C	C	.	C	.	T	T	G	.	.	A	A	.	T	G	3
H7 MeEs	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2
H8 SqSh	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
H9 IWD041J	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	T	G	T	.	.	.	.	C	G	3
H10 KD302J	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	T	G	T	G	.	.	.	C	A	2
H11 TYD083J	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	T	G	.	G	.	.	.	T	G	5
H12 NND454J	.	G	G	.	.	.	.	.	.	.	.	.	.	.	.	T	G	.	.	.	A	.	C	G	1
H13 CM95, Fuka4-4	.	G	G	.	.	.	.	C	.	.	C	.	.	.	.	T	G	.	.	.	A	A	C	G	4
H14 Co-46	.	G	.	.	.	.	.	.	.	.	C	.	.	.	.	T	G	.	G	.	.	.	C	G	2
H15 MS5655J	.	G	G	C	.	.	.	.	.	.	.	.	.	C	T	T	G	.	.	A	A	A	C	G	5
H16 NRB648J	.	G	.	.	.	.	.	.	.	.	.	T	.	.	.	T	G	.	.	.	.	.	C	G	2
H17 MD965J	.	G	.	.	.	G	.	.	.	.	.	.	.	.	.	T	G	.	.	.	.	.	C	G	5
H18 OD291J	.	G	.	.	.	.	.	.	.	.	.	T	.	.	.	T	G	.	G	.	.	.	C	G	2
H19 OGI11X2, OGI12J3	G	G	.	.	T	.	.	.	T	C	.	T	.	.	.	T	G	.	G	.	.	.	C	G	8
H20 GFD560J	G	G	.	.	.	.	A	.	.	.	.	T	.	.	.	T	G	.	G	.	.	.	T	G	6
H21 TUR4	.	G	G	.	.	.	.	C	.	.	C	.	C	C	T	T	G	.	.	A	.	A	C	G	3
H22 TUR84	.	G	G	.	.	.	.	C	.	.	C	.	C	C	T	T	G	.	.	A	.	A	C	G	4
H23 TUR83, TUR86, TUR89	.	G	G	.	.	.	.	C	.	.	C	.	C	C	T	T	G	.	.	A	.	A	C	G	3
H24 Can	.	G	G	C	.	.	.	C	.	.	C	.	.	.	.	T	G	.	.	.	A	A	C	G	3
H25 Mf	.	G	G	.	.	.	A	.	.	.	C	.	.	.	T	T	G	.	.	A	.	.	C	G	3
H26 Va	.	G	.	.	.	.	.	.	.	.	.	.	.	C	.	T	G	.	.	.	.	.	.	G	4
H27 Z1	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	T	G	.	G	.	.	.	.	G	1
Total:																									955

Coloured boxes are used to distinguish between isolates from different countries (Blue: Iran; Yellow: Japan; Green: Turkey; Red: South Korea).

ISSN (print): 1823-8262, ISSN (online): 2231-7538

Malays. J. Microbiol. Vol 18(1) 2022, pp. 79-92
DOI: http://dx.doi.org/10.21161/mjm.211276

Table 4: The country-specific variation sites (marked in red) determined via the haplotype analysis.

Isolate	Country	Variation site (bp)	Nucleotide substitution	Change in amino acid coded	Type of variation
IRN-BRE5	Iran	93	TCC → CCC	S → P	Non- synonymous
IWD401J	Japan	85	GCG → TCG	A → S	Non-synonymous
TUR83	Turkey	489	GCT → GCC	A → A	Synonymous
Z1	South Korea	-	-	-	-

Table 5: The general coat protein profiles of the selected four CMV isolates.

Isolate	Amino acid length	Molecular weight (Da)	Theoretical pI	Subcellular localization	Top five structural neighbors
IRE-BRE5	218	24114.53	9.92	Outer membrane	1f15-C (coat protein) 1js9-C (coat protein) 4y6t-E (coat protein) 4nwv-A (coat protein) 1bmv-2 (movement protein)
IWD401J	218	24104.49	9.92	Outer membrane	1f15-C (coat protein) 1js9-C (coat protein) 4y6t-E (coat protein) 1bmv-2 (movement protein) 4nwv-A (coat protein)
TUR83	218	24112.56	9.92	Outer membrane	1f15-C (coat protein) 1js9-C (coat protein) 4y6t-E (coat protein) 4nwv-A (coat protein) 1bmv-2 (movement protein)
Z1	218	24114.53	9.92	Outer membrane	1f15-C (coat protein) 1js9-C (coat protein) 4y6t-E (coat protein) 4nwv-A (coat protein) 1bmv-2 (movement protein)

to Iran isolates (except for Iran IRN-TVRa26 and Japan MS56555J). Site 489 was Turkey-specific (except for Iran IRN-Tlm1). These country-specific SNP markers are essential and potential biogeographical genetic markers to trace the country of origin of an unknown new CMV isolates. There are no South Korea specific SNP detected in this study and this has been previously proven by Kim *et al.* (2014), they postulated that biogeographical migration of CMV isolates from other countries had contributed towards the loss of South Korea isolate uniqueness.

The identified country-specific SNP markers were subjected to amino acid variation analysis to determine whether the variation is synonymous or non-synonymous (Table 4). Interestingly, the Iran and Japan-specific SNP markers have resulted in a non-synonymous variation in the amino acid sequences. The cytosine nucleotide specific to Iran isolates had yielded proline in place of serine in other isolates. The serine in Japan isolates was produced in place of alanine, as the result of the Japan-specific thymine. Conversely, the nucleotide variation spotted between Turkey and other country isolates did not lead to an amino acid change, thus a synonymous variation was detected. Further inspection on the predicted coat protein structure unraveled that both Site 85 and 93 are not involved in the formation of essential protein conformation. A similar outcome was observed in piper yellow mottle virus when global reverse transcriptase gene comparison was performed by Lim *et al.* (unpublished) where country-specific SNP markers do not interfere with the protein secondary structure conformation.

Coat protein modelling and gene ontology prediction depicted high conservation across CMV isolates from different countries

The general protein profiles were tabulated in Table 5, displaying the amino acid lengths, molecular weights, theoretical isoelectric points (pI), subcellular localizations as well as top five structural neighbors of the CMV coat proteins examined in this study. The amino acid lengths and theoretical pIs are highly conserved for all isolates. The molecular weights, however, did not differ more than 3 kDa with the South Korea Z1 and Iran IRE-BRE5 isolates scoring the highest (24114.53 kDa) whereas the Turkey TUR83 scored the lowest (24112.56 kDa). All coat proteins were predicted to localize at the outer membrane. The top five structural neighbors of all isolates do not differ, where 1f15-C, 1js9-C, 4y6t-E and 4nmv-A all refer to coat protein and 1bmv-2 refers to movement protein.

 ISSN (print): 1823-8262, ISSN (online): 2231-7538

Malays. J. Microbiol. Vol 18(1) 2022, pp. 79-92
DOI: http://dx.doi.org/10.21161/mjm.211276

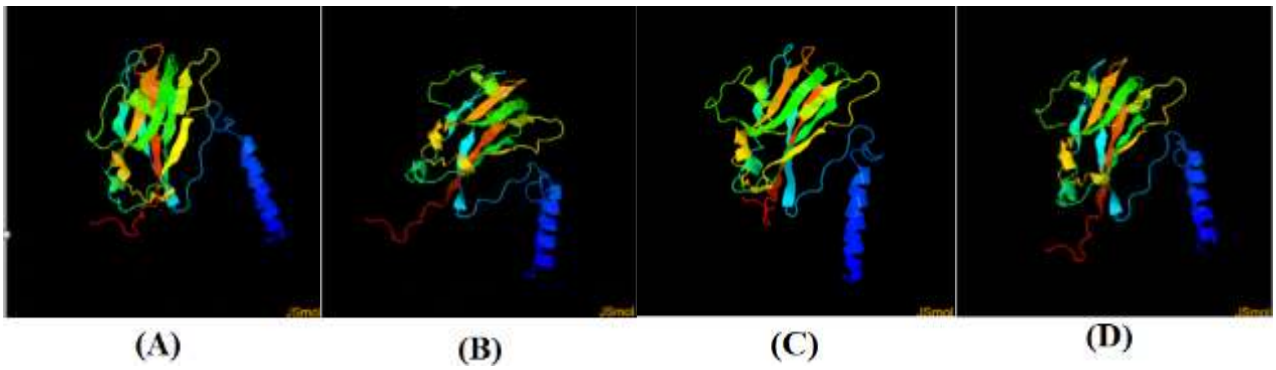
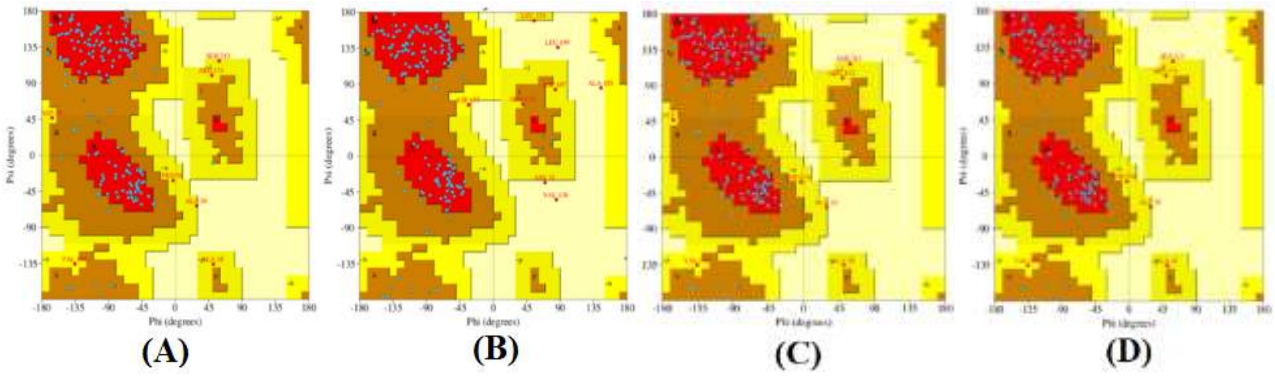


Figure 5: The three-dimensional protein modelling of coat proteins of (A) IRN-BRE5, (B) IWN401J, (C) TUR83 and (D) Z1 isolates.



Isolate	Number of residues in most favoured region	Number of residues in additional allowed region	Number of residues in generously allowed region	Number of residues in disallowed regions
IRN-BRE5	134	28	6	0
IWD401J	128	31	3	5
TUR83	134	28	6	0
Z1	134	28	5	0

Figure 6: The Ramachandran plots for coat proteins of (A) IRN-BRE5, (B) IWN401J, (C) TUR83 and (D) Z1 isolates.

The three-dimensional protein modelling further verified the high conservation in secondary protein structure across all four CMV coat proteins from four different countries (Figure 5). All four coat proteins have five alpha helices and nine anti-parallel beta sheets. The predicted protein structures were quality checked with the plotting of the Ramachandran plots (Figure 6). Only the Japan IWD401J isolate has five residues located in the disallowed regions. This phenomenon may be contributed by the slight nucleotide change induced by the Japan-specific SNP marker, which had resulted in the non-synonymous amino acid change without causing a disruption in the protein structure.

The gene ontology analysis of CMV coat proteins across countries like Turkey, Japan, South Korea and Iran unraveled an analogous spectrum of pattern (Figure 7). The cellular component encompasses most of the high score gene ontology terms across all four isolates. The binding gene ontology term scored close to that of the terms from the cellular component group. The gene ontology terms under the biological processes category displayed a diverse pattern in contrast to other group across all four isolates. Interestingly, the Iran IRN-BRE5 scored the same hits and score values as South Korea Z1for all three gene ontology categories, except for the entry hydrolase activity in IRN-BRE5 and nucleotide binding in Z1. This outcome suggest that similar functionality is probably shared by both Iran and South Korea isolates coat proteins.

 ISSN (print): 1823-8262, ISSN (online): 2231-7538

Malays. J. Microbiol. Vol 18(1) 2022, pp. 79-92
DOI: http://dx.doi.org/10.21161/mjm.211276

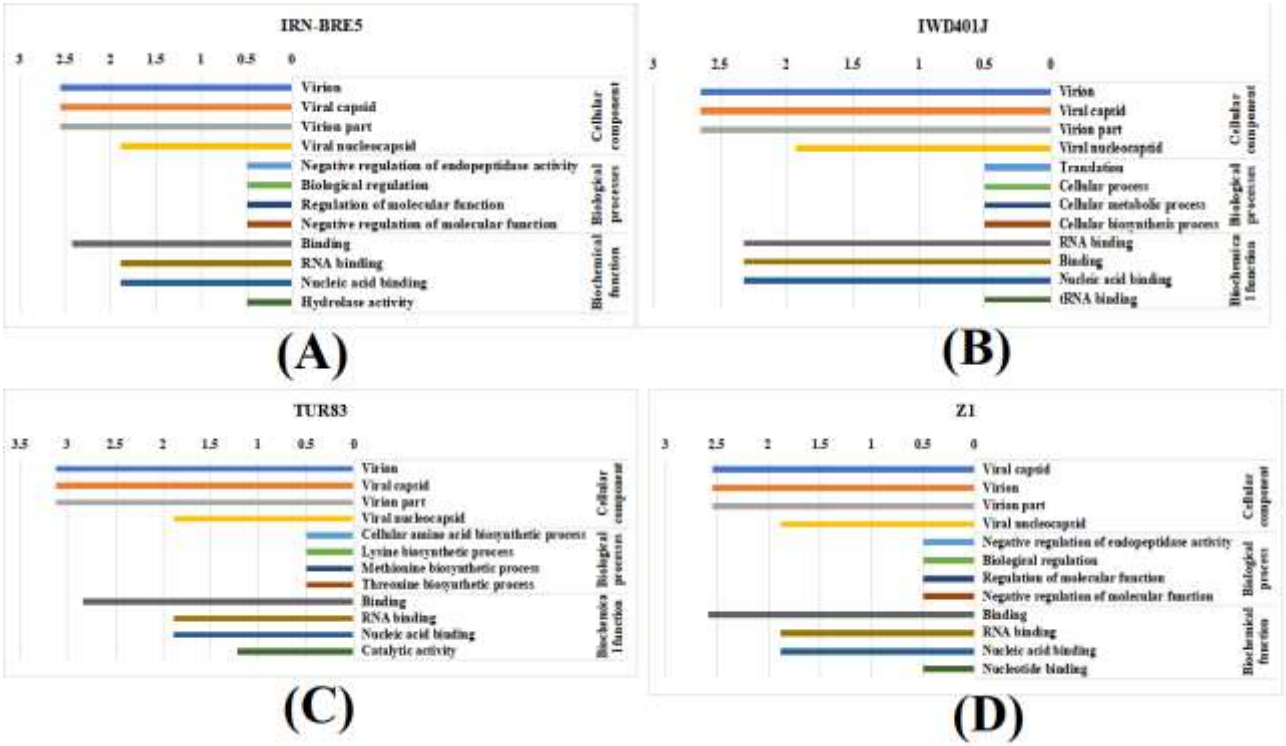


Figure 7: The gene ontology predictions for coat proteins of (A) IRN-BRE5, (B) IWN401J, (C) TUR83 and (D) Z1 isolates.

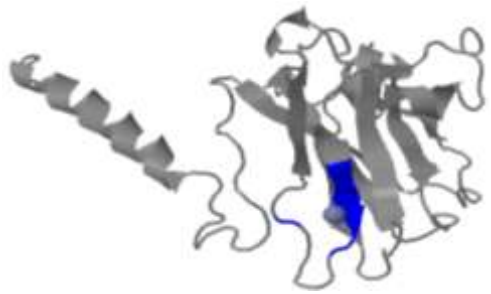


Figure 8: The predicted binding sites of IRN-BRE5 coat protein. The binding site is colored blue and the gray sphere represents zinc heterogen.

The binding sites of the coat proteins are similar across all four isolates investigated, namely valine152, glutamine154, glutamine161, tyrosine153 and tyrosine155 (Figure 8). These sites are essential starting points for the examination of the association between virulence and binding affinity of the virus in the future. Besides, a zinc heterogen was observed at the binding site of the coat protein. This element is vital in the stabilization of negative charge when CMV binds to plant cells as a result of the strong Lewis acid characteristic of the zinc molecule (Kambe *et al.*, 2015). Moreover, the activation of membrane fusion across CMV and target plant cells for genetic material transfer necessitates zinc heterogen as one of the major supplements (York and Nunberg, 2007).

Future perspectives

The success in the search for the association between single nucleotide polymorphic markers, pathogenicity and virulence was observed in many plant viruses and pathogens such as the rice blast fungus, tomato virus, chestnut pathogen, cassava brown streak virus, mungbean yellow mosaic virus and sugarcane leaf virus (Yadav *et al.*, 2015; Xu *et al.*, 2017; Ganeshan *et al.*, 2018; Oyesigye *et al.*, 2018; Pimenta *et al.*, 2021; Stauber *et al.*, 2021). One of the most studied target is the rice blast fungus where many researchers have identified markers of virulence via genome-wide association studies. This genome-wide approach has been proven powerful in the discovery of novel gene functions related to rice blast disease previously unachievable using conventional experimental analysis (Ganeshan *et al.*, 2018). The identified SNP markers were very effective in indicating the vital pathogenicity and virulence loci associated with overcoming resistance found in isolates harvested from a particular country of origin and host species (Ganeshan *et al.*, 2018). This research pipeline can be implemented on the cucumber mosaic virus to further enhance our comprehension on the CMV pathogenicity and virulence associated gene loci in the future. With the data generated from this study, the direction for the genome-wide association studies will be much clearer and straightforward.

 ISSN (print): 1823-8262, ISSN (online): 2231-7538

CONCLUSION

In a nutshell, the global genome comparison of CMV in this study had yielded imperative data previously undiscovered. One major finding of this study is that the identified country-specific single nucleotide polymorphic markers are essential for future global characterization of the virulence and pathogenicity of CMV in future. For instance, the genome size can be a potential preliminary country-specific marker for South Korea and the GC content can be used to preliminarily distinguish Turkey isolates from the others. The motif analysis as well as whole genome and coat protein phylogenetic trees failed to resolve country-specific clusters. Further coat protein haplotype analysis revealed country-specific SNP markers for Iran, Turkey and Japan isolates. Lastly, coat protein modelling and gene ontology prediction depicted high conservation across CMV isolates from different countries. It is hoped that the endeavors to unravel the virulence and pathogenicity of CMV could be initiated as soon as possible via genome-wide association studies since the genome availability is quite abundant and sufficient. By then, the harvest loss caused by this virus can be greatly reduced via the most effective biological control measures in the future.

ACKNOWLEDGEMENTS

This research was fully supported by Fundamental Research Grant Scheme with grant number FRGS/1/2019/STG05/UNIMAS/03/1 awarded to H. H. Chung by The Ministry of Higher Education Malaysia. The author would like to thank Universiti Malaysia Sarawak on the support of this publication.

AUTHOR'S CONTRIBUTION

Conceptualization: Leonard Whye Kit Lim, Hung Hui Chung; Methodology: Leonard Whye Kit Lim, Ing Mee Hung; Formal analysis and investigation: Leonard Whye Kit Lim, Ing Mee Hung; Writing - original draft preparation: Leonard Whye Kit Lim, Ing Mee Hung; Writing - review and editing: Hung Hui Chung; Funding acquisition: Hung Hui Chung; Resources: Hung Hui Chung; Supervision: Hung Hui Chung.

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare that are relevant to the content of this article.

REFERENCES

Bailey, T. L., Johnson, J., Grant, C. E. and Noble, W. S. (2015). The MEME Suite. *Nucleic Acid Research* **43(W1), W39-W49.**

de Castro, E., Sigrist, C. J. A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., Bairoch, A. and Hulo, N. (2006). ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research* **34, W362-W365.**

Draeger, K. R. (2016). Cucumber mosaic. University of Wisconsin-Madison: https: //hort. extension.wisc.edu/articles/cucumber-mosaic/ [Retrieved on 24 April 2021].

Frazer, K. A., Pachter, L. Poliakov, A., Rubin, E. M. and Dubchak, I. (2004). VISTA: Computational tools for comparative genomics. *Nucleic acid Research* **32, W273-W279.**

Ganeshan, V. D., Opiyo, S. O., Mutiga, S. K., Rotich, F., Thuranira, D. M., Were, V. M., OuéDraogo, I., Zhou, B. O., Soanes, D. M., Correll, J. C., Wang, G. L., Talbot, N. J. and Mitchell, T. K. (2018). A genome-wide association study identifies SNP markers for virulence in *Magnaporthe oryzae* isolates from sub-saharan Africa. bioRxiv: https://www.biorxiv.org/content/10.1101/418509v1.full

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D. and Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server. *In*: The Proteomics Protocols Handbook. Walker, J. M. (ed.). Humana Press, New Jersey. **pp. 571-607.**

GenBank. (2021). Cucumber mosaic virus. https://www.ncbi.nlm.nih.gov/nuccore/?term=cucumber+mosaic+virus [Retrieved on 16 September 2021].

Grant, J. R., Arantes, A. S. and Stothard, P. (2012). Comparing thousands of circular genomes using the CGView comparison tool. *BMC Genomics* **13, 202.**

Holm, L., Kaarianen, S., Rosenstrom, P. and Schenkel, A. (2008). Searching protein structure database with DaliLite v.3. *Bioinformatics* **24(23), 2780-2781.**

Jacquemond, M. (2012). Cucumber mosaic virus. *Advances in Virus Research* **84, 439-504.**

Kambe, T., Tsuji, T., Hashimoto, A. and Itsumura, N. (2015). The physiological, biochemical, and molecular roles of zinc transporters in zinc homeostasis and metabolism. *Physiological Reviews* **95(3), 749-784.**

Katsir, L., Zhepu, R., Garcia, D. S., Piasezky, A., Jiang, J., Sela, N., Freilich, S. and Bahar, O. (2018). Genome analysis of haplotype D of *Candidatus* Liberibacter solanacearum. *Frontiers in Microbiology* **9, 2933.**

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. and Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modelling, prediction and analysis. *Nature Protocols* **10(6), 845-858.**

Kim, M. K., Seo, J. K., Kwak, H. R., Kim, J. S., Kim, K. H., Cha, B. J. and Choi, H. S. (2014). Molecular genetic analysis of cucumber mosaic virus populations infecting pepper suggests unique patterns of evolution in Korea. *The American Phytopathological Society* **104(9), 993-1000.**

Kimaru, S. L., Kilalo, D. C., Muiru, W. M., Kimenju, J. W. and Thuku, C. R. (2020). Molecular detection of cucumber mosaic virus and tabacco mosaic virus infecting African nightshades (*Solanum scabrum*

Miller). *Internationl Journal of Agronomy* **2020, Article ID 8864499.**

Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K. (2018). MEGA X: Molecular evolutionary genetic analysis across computing platforms. *Molecular Biology and Evolution* **35(6), 1547-1549.**

Lim, L. W. K., Chung, H. H. and Hussain, H. (2020a). Complete chloroplast genome sequencing of sago palm (*Metroxylon sagu* Rottb.): Molecular structures, comparative analysis and evolutionary significance. *Gene Reports* **19, 100662.**

Lim, L. W. K., Chung, H. H. and Hussain, H. (2020b). Organellar genome copy number variations and integrity across different organs, growth stages, phenotypes and main localities of sago palm (*Metroxylon sagu* Rottboll) in Sarawak, Malaysia. *Gene Reports* **21, 100808.**

Lim, L. W. K., Chung, H. H., Chong, Y. L. and Lee, N. K. (2018). A survey of recently emerged genome-wide computational enhancer predictor tools. *Computational Biology and Chemistry* **74, 132-141.**

Lim, L. W. K., Chung, H. H., Lau, M. M. L., Aziz, F. and Gan, H. M. (2021). Improving the phylogenetic resolution of Malaysian and Javan mahseer (Cyprinidae), *Tor tambroides* and *Tor tambra*: Whole mitogenomes sequencing, phylogeny and potential mitogenome markers. *Gene* **791, 145708.**

Lim, L. W. K., Liew, J. X. and Chung, H. H. (unpublished). Piper yellow mottle virus: A deep dive into the genome.

Lin, Y., Hu, F., Tang, J. and Moret, B. M. E. (2013). Maximum likelihood phylogenetic reconstruction from high-resolution whole-genome data and a tree of 68 eukaryotes*. Pacific Symposium on Biocomputing* **2013, 285-296.**

Nematollahi, S., Sokhandan-Bashir, N., Rakhshandehroo, F. and Zamanizadeh, H. R. (2012). Phylogenetic analysis of new isolates of cucumber mosaic virus from Iran on the basis of different genomic regions. *The Plant Pathology Journal* **28(4), 381-389.**

Ng, J. C., Liu, S. and Perry, K. L. (2000). Cucumber mosaic virus mutants with altered physical properties and defective in aphid vector transmission. *Virology* **276, 395-403.**

Nouri, S., Arevalo, R., Falk, B. W. and Groves, R. L. (2014). Genetic structure and molecular variability of cucumber mosaic virus isolates in the United States. *PLoS ONE* **9(5), e96582.**

Oyesigye, E., Zacarias, A., Mondjana, A., Magaia, H. and Ferguson, M. (2018). Single nucleotide polymorphism (SNP) diversity of cassava genotypes in relation to cassava brown streak disease in Mozambique. *Plant Genetic Resources* **16(6), 533-543.**

Pimenta, R. J. G., Aono, A. H., Burbano, R. C. V., Coutinho, A. E., da Silva, C. C., dos Anjos, I. A., Perecin, D., de Andrade Landell, M. G., Goncalves, M. C., Pinto, L. R. and de Souza, A. P. (2021). Genome-wide approaches for the identification of markers and genes associated with sugarcane yellow leaf virus resistance. *Scientific Reports* **11, 15730.**

Qiu, Y., Zhang, Y., Wang, C., Lei, R., Wu, Y., Li, X. and Zhu, S. (2018). Cucumber mosaic virus coat protein induces the development of chlorotic symptoms through interacting with the chloroplast ferredoxin I protein. *Scientific Reports* **8, 1205.**

Shakya, M., Ahmed, S. A., Davenport, K. W., Flynn, M. C., Lo, C. C. and Chain, P. S. G. (2020). Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. *Scientific Reports* **10, 1723.**

Stauber, L., Badet, T., Feurtey, A., Prospero, S. and Croll, D. (2021). Emergence and diversification of a highly invasive chestnut pathogen lineage across southeastern Europe. *eLife* **10, e56279.**

Wass, M. N., Kelley, L. A. and Sternberg, M. J. (2021). 3DLigandSite: Predicting ligand-binding sites using similar structures. *Nucleic Acids Research* **38, W469-W473.**

Xu, C., Sun, X., Taylor, A., Jiao, C., Xu, Y., Cai, X., Wang, X., Ge, C., Pan, G., Wang, Q., Fei, Z. and Wang, Q. (2017). Diversity, distribution, and evolution of tomato viruses in China uncovered by small RNA sequencing. *Journal of Virology* **91(11), e00173-17.**

Yadav, C. B., Bhareti, P., Muthamilarasan, M., Mukherjee, M., Khan, Y., Rathi, P. and Prasad, M. (2015). Genome-wide SNP identification and characterization in two soybean cultivars with contrasting mungbean yellow mosaic India virus disease resistance traits. *PLoS ONE* **10(4), e0123897.**

Yang, M., Derbyshire, M. K., Yamashita, R. A. and Marchler-Bauer, A. (2020). NCBI's conserved domain database and tools for protein domain analysis. *Current Protocols in Bioinformatics* **69(1), e90.**

Yokono, M., Satoh, S. and Tanaka, A. (2018). Comparative analyses of whole-genome protein sequences from multiple organisms. *Scientific Reports* **8, 6800.**

York, J. and Nunberg, J. H. (2007). A novel zinc-binding domain is essential for the formation of the functional Junin virus envelope glycoprotein complex. *Journal of Virology* **81(24), 13385-13391.**

Zhang, X. P., Liu, D. S., Yan, T., Fang, X. D., Dong, K., Xu, J., Wang, Y., Yu, J. L. and Wang, X. B. (2017). Cucumber mosaic virus coat protein modulates the accumulation of 2b protein and antiviral silencing that causes symptom recovery *in planta*. *PLoS Pathogens* **13(7), e1006522.**

Zitter, T. A. and Murphy, J. F. (2009). Cucumber mosaic virus. *The Plant Health Instructor*. doi: 10.1094/PHI-I-2009-0518-01